Statistical and Computational Methodology for the Analysis of Forensic DNA Mixtures with Artefacts

Therese Graversen
Jesus College, Oxford

Hilary Term 2014

A thesis submitted for the degree of Doctor of Philosophy in Statistics

Abstract

This thesis proposes and discusses a statistical model for interpreting forensic DNA mixtures. We develop methods for estimation of model parameters and assessing the uncertainty of the estimated quantities. Further, we discuss how to interpret the mixture in terms of predicting the set of contributors.

We emphasise the importance of challenging any interpretation of a particular mixture, and for this purpose we develop a set of diagnostic tools that can be used in assessing the adequacy of the model to the data at hand as well as in a systematic validation of the model on experimental data.

An important feature of this work is that all methodology is developed entirely within the framework of the adopted model, ensuring a transparent and consistent analysis.

To overcome the challenge that lies in handling the large state space for DNA profiles, we propose a representation of a genotype that exhibits a Markov structure. Further, we develop methods for efficient and exact computation in a Bayesian network. An implementation of the model and methodology is available through the R package DNAmixtures.

Acknowledgements

I would like to thank my supervisor Steffen Lauritzen, whose honest excitement about the field of DNA mixture analysis has been a great inspiration. Our collaboration so far has been enjoyable and highly productive. I am indebted to Robert Cowell, Julia Mortera, and Steffen Lauritzen for letting me join their quest for statistical methodology for DNA mixture analysis. Financially, this has only been possible due to funding from the Department of Statistics and the EPSRC.

The development of my software DNAmixtures has been a long journey so far. A large part of DNAmixtures has been written during a stay at the Statistical Laboratory, Cambridge, for which I am grateful to Phillip Dawid. Julia Mortera and Steffen Lauritzen have patiently been exposed to the multiple versions of DNAmixtures, giving me invaluable feedback throughout the development. The swiftness with which Kjell Konis has dealt with my long and technical issues – sometimes involving major changes to RHugin – is admirable. Likewise have I had many long and detailed replies from the group behind HUGIN in my various battles with Bayesian networks of immense sizes.

I would also like to thank Marjan Sjerps from the Netherlands Forensic Institute (NFI) for providing a dataset that could be discussed openly. I have benefited hugely from visiting the NFI and I am grateful for the support, encouragement, and late night banter provided by especially Marjan Sjerps, Ate Kloosterman, Klaas Slooten, and Jerien Koopman.

I have been lucky to have the opportunity to think out loud in the company of Arnaud Doucet, Michael Sørensen, and Rasmus L. Petersen, which has had valuable influence on my work. A huge thank you is due to Luke Kelly, who has many times generously provided me with a place to stay. Finally, I would like to thank again Rasmus L. Petersen, whom I can always trust to be by my side.

Therese Graversen

Contents

Preface **1**

1 Introduction to DNA mixture analysis **9**

1.1 DNA profiling using STR marker systems 9

1.1.1 DNA profiles 10

1.1.2 The polymerase chain reaction process 12

1.1.3 Capillary electrophoresis 15

1.1.4 Artefacts 18

1.2 DNA mixture analysis 20

1.2.1 Weight of evidence 21

1.2.2 Mixture deconvolution 22

1.2.3 Evidential efficiency 23

1.3 An example case: MC15 and MC18 27

2 A statistical model for DNA mixtures **31**

2.1 Statistical model for DNA profiles 31

2.1.1 Allele counts and allele frequencies 32

2.1.2 F_{ST} and sampling-adjustment parameters 33

2.2 Peak-height distribution for DNA samples of known composition . . . 33

2.2.1 The gamma model 34

2.2.2 Stutter 40

2.2.3 Dropout 43

2.2.4 Dropin 46
2.2.5 Summary 46
2.3 Likelihood function 48
2.4 A joint model for multiple mixtures 49
2.5 A qualitative model 51

3 Bayesian network techniques **53**
3.1 An introduction to inference in Bayesian networks 54
3.1.1 Bayesian network representation 54
3.1.2 Junction tree representation 55
3.1.3 Local computations on a junction tree 61
3.2 Expectations of products using auxiliary variables 64
3.2.1 Summation by propagation 64
3.2.2 Constructing auxiliary variables 65

4 A Bayesian network representation of the DNA mixture model **69**
4.1 Markov representation of a genotype 70
4.2 Auxiliary variables for evaluation of the likelihood function 72
4.3 Posterior distributions of genotypes 75
4.3.1 Simulation under the model 76
4.4 Junction-tree representations and complexity considerations 77
4.4.1 Markov genotype representation 77
4.4.2 Alternative genotype representations 84
4.5 Extensions and modifications 85
4.5.1 Representation of multiple mixtures 85
4.5.2 Representation of the qualitative model 85
4.5.3 Amelogenin 86
4.5.4 Silent alleles 87

5 Statistical interpretation of a DNA mixture **89**

5.1 Maximum-likelihood estimation 90

5.2 Weight of Evidence 98

5.2.1 Quantitative peak information 99

5.2.2 Qualitative peak information 102

5.3 Prediction of DNA profiles 104

5.3.1 Identifying highest-probability sets 106

5.3.2 Summarising the DNA profile distribution 109

5.4 Identifying stutter and dropout 110

6 Challenging the interpretation of a mixture **113**

6.1 Uncertainty of estimated quantities 113

6.1.1 Simulated MLE 114

6.1.2 Asymptotic variance for the MLE 119

6.2 Hypothesis testing 126

6.2.1 Assumptions on parameters 126

6.2.2 The number of contributors 127

6.3 Assessing the peak height distributions 132

6.3.1 The cumulative distribution function 135

6.3.2 Assessing the distribution of above-threshold peaks 140

6.3.3 Checking for trends in the EPG 142

6.4 Prequential monitoring of peak presence 147

7 The DNAmixtures package **153**

7.1 The `DNAmixture` 154

7.1.1 The Bayesian network representation 158

7.1.2 Conditional distributions 160

7.1.3 The likelihood function 161

7.2 Maximum likelihood estimation 162

7.2.1 Imposing constraints on the parameters . . . 162
7.2.2 Asymptotic variance . . . 164
7.3 Prediction . . . 166
7.3.1 Fitted probabilities . . . 166
7.3.2 Highest probability sets of genotypes . . . 167
7.4 Simulation . . . 169
7.5 Graphics . . . 170
7.5.1 Electropherograms . . . 170
7.5.2 Diagnostic plots . . . 172

8 A case analysis: Four replicates of a low-template single-source DNA sample 177
8.1 Single profile analyses . . . 179
8.2 Joint analyses . . . 182
8.3 Unexplained artefacts . . . 185

9 Discussion 189
9.1 Assumptions about peak height distributions for fixed genotypes . . . 190
9.1.1 Marker- and allele-dependent parameters . . . 190
9.1.2 Other models using the gamma distribution . . . 191
9.1.3 Branching processes and PCR . . . 195
9.2 Assumptions about genotype distributions . . . 203
9.3 Handling unknown parameters . . . 205
9.3.1 Alternative optimisation methods . . . 205
9.3.2 Alternative estimation methods . . . 206
9.4 Final remarks . . . 208

A Data 209
A.1 MC15 and MC18 . . . 209
A.2 Low-template DNA from the NFI . . . 210

Preface

During my doctoral studies, I have joined the long-standing collaboration between Robert Cowell, Julia Mortera, and Steffen Lauritzen concerning statistical analysis of forensic DNA mixtures – DNA samples containing DNA from multiple individuals. This has resulted in a new statistical model presented in Cowell et al. (2015).

I have managed to extend their efforts in two main directions. One is concerned with the further inference in the model, such as estimation and methodology for model checking. The second direction concerns the practical implementation of the model framework, which has required the development of suitable computational methodology. In particular, a full implementation of the model of Cowell et al. (2015) and the developed statistical methodology has been made available as an R package:

DNAmixtures: Statistical analysis of mixed traces of DNA with artefacts

Therese Graversen.

R package version 0.1-3. 2014.

`http://dnamixtures.r-forge.r-project.org`.

Further, a large part of the work presented in this thesis has been made available in the form of three papers.

Estimation of Parameters in DNA Mixture Analysis

Therese Graversen and Steffen Lauritzen.

Journal of Applied Statistics, 40(11):2423–2436. 2014.

Analysis of Forensic DNA mixtures with Artefacts

Robert Cowell, Therese Graversen, Steffen Lauritzen, and Julia Mortera.

Read before the Royal Statistical Society on 11 June 2014. To appear with discussion in Journal of the Royal Statistical Society, Series C. 2015.

Computational Aspects of DNA Mixture Analysis

Therese Graversen and Steffen Lauritzen.

Statistics and Computing, DOI:10.1007/s11222-014-9451-7. 2014.

In the following, I give a brief overview of the topics covered in this thesis as well as their relation to the above publications, before returning to an in-depth introduction to DNA mixture analysis in Chapter 1.

Statistical analysis of DNA mixtures

In the setting of forensic identification using DNA, a person is identified via a DNA profile, which summarises the individual's DNA in a set of so-called genotypes. Each genotype consists of a pair of alleles that are sub-sequences of the individual's DNA.

The presence of individual alleles in a DNA sample is represented by a peak in an electropherogram (EPG), where the height of the peak is proportional to the total amount of the allele in the sample.

Inference about the underlying individual DNA profiles is done statistically, using the information in the EPG. The presence of artefacts such as dropout (the allele is present in the mixture, but no peak is detected) and stutter (the allele is not present, but a peak is detected) further complicates such an analysis.

A statistical model. Building on an earlier series of papers (Cowell et al., 2007a, 2011), in Cowell et al. (2015) we propose and discuss a statistical model for DNA mixtures in the potential presence of various complicating factors. We demonstrate that the model is flexible enough to allow the computation of quantities relevant to

crime case analysis, yet it is simple enough to render computations feasible, even for a mixture of a higher number of individuals (up to 6) than normally discussed in the literature (2-3).

Estimation. In Graversen and Lauritzen (2013), parameters in the basic gamma model (Cowell et al., 2007a) were estimated by maximum likelihood using the data from only the case at hand. It was also demonstrated how, if available, prior information on the parameters can be included and used either for penalised maximum likelihood estimation or in an MCMC sampling scheme. For the model with artefacts (Cowell et al., 2015) we estimate by the method of maximum likelihood, and we assess the uncertainty about the parameters through an estimated asymptotic variance matrix based on the numerically derived Hessian of the log-likelihood function.

Challenging the interpretation of a mixture. In practical case analysis little attention is devoted to checking that the adopted model suitably explains the case at hand. This may partly be due to a tradition of applying models that have been *validated* on large sets of experimental data, and partly due to unavailability of suitable model checking methods. We propose in Graversen and Lauritzen (2014) a suite of diagnostic tools that can assist in assessing whether the hypothesised explanation of a DNA mixture is indeed adequate; we have in this connection also found the diagnostic tools to be valuable in detecting errors in the data. Another important application of these methods would be in a systematic validation of the model on experimental data.

Exact computations using Bayesian network techniques

An important task is to ensure that the model is feasible to work with in practice.

A hypothesis offers an explanation of the composition of the mixture, in particular in terms of a set of contributors to the mixture. The DNA profiles for the contributors are given a distribution; the DNA profiles of profiled individuals are considered as fixed and known, whereas the DNA profiles of unprofiled – unknown – contributors are considered mutually independent and sampled from a suitable reference population. Models for peak heights are generally of the form

$$\Pr(E \mid H) = \sum_{\mathbf{g}} \Pr(E \mid \mathbf{g}) \Pr(\mathbf{g} \mid H)$$

with $\mathbf{g}$ denoting the DNA profiles of all contributors in the hypothesis H under consideration, and where the distribution $\Pr(E \mid \mathbf{g})$ of peak heights is conditionally independent of the hypothesis given the DNA profiles.

The size of the space of possible DNA profiles $\boldsymbol{g}$ renders it infeasible to compute this sum directly. Thus, a challenge in modelling DNA mixtures lies in handling the enormous state space of DNA profiles for unknown contributors, in particular as the number of contributors increases.

Bayesian network representation of a genotype. Cowell et al. (2007a, 2011) as well as Graversen and Lauritzen (2013) used a compact representation of a genotype by the unordered pair of two alleles, which in practice limits analysis to 2-3 contributors. Graversen and Lauritzen (2014) propose to represent a genotype by the vector of allele counts, counting for each allele how many of this type that the contributor possesses. Markov properties of the vector of allele counts then allow an efficient representation in terms of a Bayesian network, on which computations may be done locally on a few alleles.

Computation by auxiliary variables in a Bayesian network. Graversen and Lauritzen (2014) give a general method for computing expectations with respect to a set of variables in a Bayesian network. The method is based on the idea that an expectation with respect to network variables can be expressed in terms of a

probability of a configuration of binary auxiliary variables, which can in turn be computed by probability propagation.

In DNA mixture analysis, many of the quantities of interest may be expressed in terms of expectations with respect to the genotypes of unknown contributors, and this is exploited in Graversen and Lauritzen (2014), where the Markov representation of a genotype is used in combination with various auxiliary variables. We demonstrate how computation by auxiliary variables allows efficient evaluation of the likelihood function; this enables, for instance, numerical maximisation of the likelihood. Further, through conditioning on the introduced auxiliary variables, we obtain a Bayesian network representation of the conditional distribution of genotypes given a set of observed peak heights. The methodology is also useful for model checking purposes, for instance by allowing the evaluation of the cumulative distribution function for peak heights.

Complexity considerations and junction trees. In practice, the size of the networks is a major concern in terms of both memory consumption and computation time for the probability propagation. The complexity is directly related to the total size of the junction tree, and in Graversen and Lauritzen (2014) we discuss for a range of junction trees the relation of the total size to the number of possible alleles at a marker as well as to the number of unknown contributors.

Our best junction tree has a total size which is linear rather than polynomial in the number of possible alleles. The total size still grows exponentially in the number of unknown contributors, however with a lower growth rate than for previous Bayesian network representations, e.g. as used in Cowell et al. (2007a, 2011). The gain in efficiency by using our best junction tree is reflected in the fact that we have been able to maximise the likelihood function for a model with six unknown contributors and that models with up to five unknown contributors can easily be handled on a standard desktop computer.

Putting it all into practice

A final topic is whether the methodology presented is useful for DNA mixture analysis.

Implementation as an R package. Our model is implemented along with tools for estimation and model checking in the R package DNAmixtures (Graversen, 2014). Implementations of models and methodology for DNA mixture analysis have only in few cases been made available to a wider audience – Steele and Balding (2014) give an overview of six currently available pieces of software, DNAmixtures being one.

The package is implemented using the computational approach of Graversen and Lauritzen (2014) described above. By providing the full implementation, we also hope to emphasise that the methodology is indeed computationally feasible to work with in practice. The package is open source, but depends on HUGIN (HUGIN API, 2009) for the computations on a Bayesian network.

Data analysis. We have applied our methodology in various settings with consistent and promising results. Unfortunately the performance on specific casework data cannot openly be discussed and the general availability of suitable datasets is limited, so for illustration purposes we here discuss our methodology in the following two settings.

Our main example concerns two mixtures, allegedly composed of three or more people, relating to a pub-brawl case discussed in Gill et al. (2008). These mixtures are extensively analysed in Cowell et al. (2015) and Graversen and Lauritzen (2014), and are also used throughout the documentation for DNAmixtures.

The Netherlands Forensic Institute (NFI) have kindly provided four replicate analyses of a low-template, single source trace, which we analyse in detail in Chapter 8. Low-template samples, where there is little DNA present, pose a challenge even in the case of a single contributor, since they are particularly prone to dropout.

Interestingly, I have been able to fully infer the DNA profile as well as detect the presence of non-explained artefacts; this demonstrates the versatility of our model.

Contributions

In summary, my specific contributions to the field have consisted of

- developing methods for efficient exact computation using Bayesian network techniques.

- developing methods for estimation of model parameters and assessing the uncertainty of the estimates.

- developing methods for diagnostic methods for assessing the adequacy of the model to the data at hand, as well as for use in a systematic validation of the model on experimental data.

- developing methods for mixture deconvolution based on the proposed model.

- making the model and methodology available to a wider audience through the R package DNAmixtures.

An important feature of this work is that all methods are developed entirely within the framework of the adopted model, ensuring a consistent analysis.

Outline

The topics of the thesis are organised as follows. After an introduction to DNA mixture analysis in Chapter 1, the new statistical model for DNA mixtures is introduced in Chapter 2. The computational aspects are discussed in Chapters 3 and 4. Methodology relating to the statistical interpretation of a DNA mixture is developed over Chapters 5 and 6. The R package DNAmixtures presented in Chapter 7 implements the model and methodology of Chapters 2-6.

Chapter 8 presents an example analysis that highlights the versatility of the model, in that it can be useful in the analysis of single-source DNA samples that are traditionally analysed by other methods than those for mixed samples.

Finally, in Chapter 9 we mention a few immediate extensions that we hope to address in the future.

Chapter 1

Introduction to DNA mixture analysis

In crime cases there is often a need for identification of people. It may, for example, be as a part of an investigative phase to identify a suspect to look for, or it may be to determine whether there is a match of a specific person to a DNA sample from the crime scene.

Here we explain some basic facts about DNA profiles and the use of DNA for forensic identification. The analysis of a sample of DNA from multiple individuals can be difficult as the detection process, which we describe below, does not directly enable the identification of the DNA profiles of the donors. In this thesis, we are particularly interested in the analysis of such mixed samples of DNA, so-called *DNA mixtures*.

1.1 DNA profiling using STR marker systems

In the following we give an overview of the DNA profiles used for forensic identification, the laboratory procedure used for analysing DNA samples, and the various artefacts that can complicate the analysis. Further details on the data collection

procedures can be found in Butler (2005) and the manuals from Applied Biosystems (Applied Biosystems, 2009, 2012b).

1.1.1 DNA profiles

A (double-stranded) DNA molecule consists of two strands with nucleobases (bases) adenine (A), thymine (T), cytosine (C), or guanine (G) attached and is usually represented by a sequence of bases using the labels A, T, C or G. The natural state of a DNA molecule is for the two strands to be linked through hydrogen bonds between the bases, and the strands align pairing bases as A–T, C–G always (Figure 1.1). A strand has an orientation; a DNA sequence is read in the direction from the *5'-end* to the *3'-end* of a strand, and two complementary strands of DNA line up in opposite direction so that the 5'-end of one strand binds to the 3'-end of the other strand. Usually only the sequence of one strand is used for the representation of the DNA molecule, as the complementary strand is completely determined by this specification.

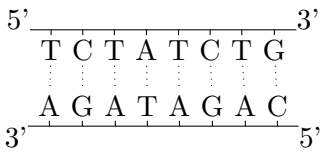


Figure 1.1: A double-stranded DNA molecule.

The DNA found in the nucleus of a cell is grouped into 23 pairs of chromosomes; 22 autosomal chromosomes and two sex-determining chromosomes. A person has for each chromosome pair one inherited from the father, and one from the mother.

Rather than identifying a person through all of the DNA, only 10-15 chosen positions on the pairs of autosomal chromosomes are considered. Each such *locus* or *marker* is chosen to exhibit variation so that distinction between people is possible.

The two DNA sequences at a marker – one for each chromosome in the pair – are called *alleles* and together they constitute the *genotype* at that marker. A *DNA profile* is the set of genotypes across a set of markers. The person is said to be

homozygote at the marker if the two alleles are the same, and *heterozygote* if they are not.

We consider DNA profiles based on *Short Tandem Repeat (STR) markers*, characterised by exhibiting variation in terms of length polymorphism; the variation between alleles lies in the number of times that a certain motif, typically consisting of four bases, is repeated. The alleles are conventionally named using an integer value for the number of repeats and a decimal value for the number of base pairs in any present incomplete repeat motif. For example, for marker TH01 the repeat structure for allele 9 is the motif AATG repeated nine times, written as $[\text{AATG}]_9$, whereas for allele 9.3 it is

$$[\text{AATG}]_6\text{ATG}[\text{AATG}]_3.$$

An allele containing incomplete repeats is called a *microvariant*. The repeat structure of an allele is surrounded by *flanking regions*, which are the same for all alleles at a locus, meaning that the position of the locus can be determined by the flanking regions; the length of the flanking regions depends on the specific kit used for the analysis. For marker D8S1179, allele 12, the repeat motif itself is $12 \times 4\text{bp} = 48\text{bp}$ long, but when including the flanking regions used with the Profiler Plus® kit, the fragment length of the allele is 143bp.

Note that the repeat structure can be complex and involve several repeated motifs; for instance, at marker D8S1179 the repeat structure for allele 12 is $[\text{TCTA}]_{12}$, whereas for allele 13 it is $[\text{TCTA}]_1[\text{TCTG}]_1[\text{TCTA}]_{11}$ rather than $[\text{TCTA}]_{13}$. Even more complex repeat structures are also possible, but for our purposes a distinction of the various structures is not important, since the measurement methods discussed in Sections 1.1.2 and 1.1.3 below only allow the detection of the length of the allele in base pairs, and not the actual sequence.

Amelogenin

The marker *Amelogenin* is used for gender identification and is situated on the sex-determining pair of chromosomes. It thus differs from the STR markers, which are chosen specifically not to code for any traits, and also the alleles for Amelogenin do not have a repeat structure. A person has either two X chromosomes (female) or one X and one Y chromosome (male). The detection of which of these chromosomes are present can be done in conjunction with the analysis of STR markers, since the two chromosomes can be distinguished by the presence of one of two possible alleles, thus denoted by X and Y. The two possible genotypes for Amelogenin are therefore XX and XY, ignoring here rare genetic variants.

1.1.2 The polymerase chain reaction process

The purpose of the *polymerase chain reaction (PCR) process* is *amplification* of the DNA sample. The PCR process is based on the replication process of DNA; it locates and repeatedly copies the DNA sequences of interest until an amount sufficient for further processing is reached.

Let us consider the amplification of the allele at a marker, starting from a single DNA molecule (Figure 1.2). The *target region* is the sequence of interest, and it consists of the repeated pattern surrounded by the two flanking regions identifying the position of the marker.

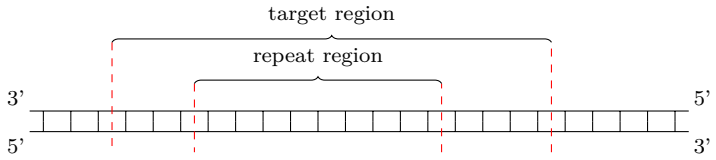


Figure 1.2: A DNA molecule with the target region corresponding to an allele.

One cycle of the PCR process consists of three steps. In the first step, the DNA strands are separated by heating up the sample to 94-96°C (Figure 1.3).

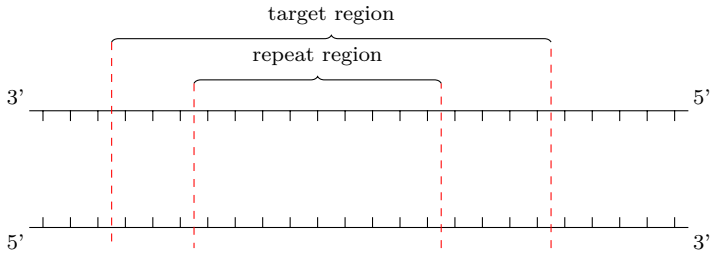


Figure 1.3: Step 1 in a PCR cycle. The strands of DNA are separated.

In the second step of the PCR cycle, the target region is localised (Figure 1.4). When the temperature is lowered to 50–65°C, a forward primer anneals to the flanking region at the 3'-end of one strand, and a reverse primer anneals to the 3'-end of the complementary strand. Typically, the forward primer has a dye attached to it.

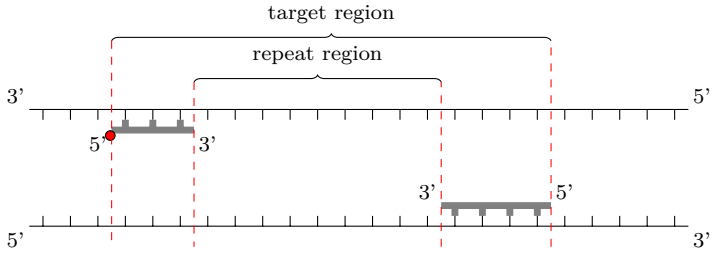


Figure 1.4: Step 2 of a PCR cycle. Primers anneal to the DNA strands.

The third step of the PCR cycle is concerned with the actual copying (Figure 1.5). When raising the temperature to 72°C, a DNA polymerase produces a strand complementary to the target sequence by extending the primers in the direction of the 5'-end, using the free floating nucleotides in the solution as building blocks.

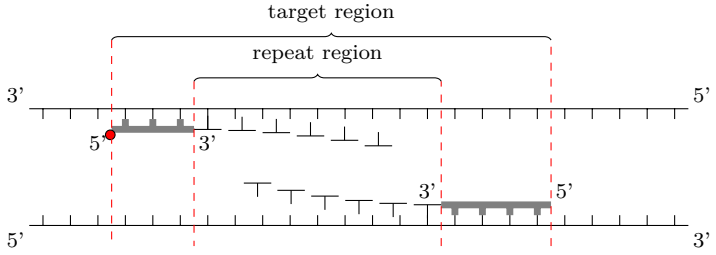


Figure 1.5: Step 3 of a PCR cycle. Primers are extended.

During the very first cycle there is nothing to mark the other flanking region and stop the polymerase from copying, and so the resulting copy will extend beyond the target region (Figure 1.6). After the first cycle we have two double-strands, each of which consists of the original strand and a PCR product that is cut off at its 5'-end. But there are no copies of the target region alone, which is what we are after.

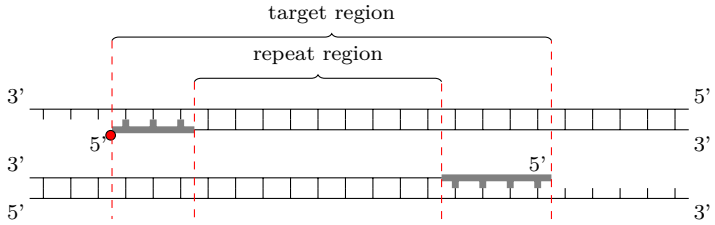


Figure 1.6: After the first cycle of PCR.

During the second cycle, consider what happens to the PCR products from the first cycle: As this new template is cut off at the 5'-end of the target region, and copying starts from the 3'-end of the target region, the resulting copy corresponds exactly to the target region (Figure 1.7).

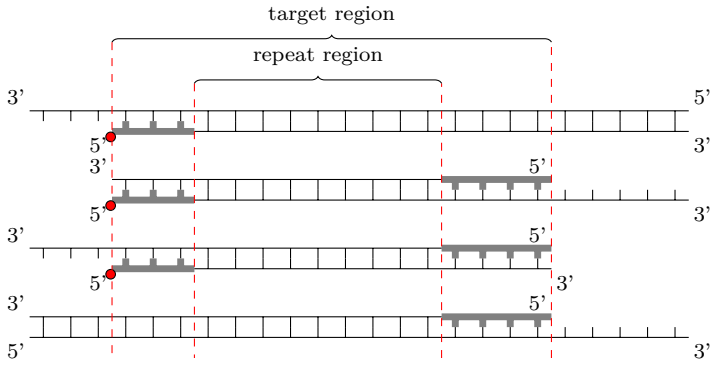


Figure 1.7: After the second cycle of PCR.

From third cycle onwards, we thus start obtaining double stranded copies of the target region alone, each with a dye attached. Note that each cycle of PCR will lead to a few copies that extend beyond the target region, since the original molecules continue to exist and produce more copies; but, as the total number of molecules grows, these constitute only a tiny proportion of the PCR product. In fact, the number of copies of the target region grows exponentially with the number

of cycles, whereas the number of original strands remains constant and the number of "half strands" grows linearly.

When the PCR process ends, we have an amplified sample containing copies of the alleles at the markers of interest in similar proportions to those of the original sample.

The PCR process has been described in various settings, for instance in Gill et al. (2005); Stolovitzky and Cecchi (1996); Sun (1995), as a branching process where the number of new copies after a cycle is binomially distributed. We briefly discuss this model in Chapter 9.1.3 below.

1.1.3 Capillary electrophoresis

The purpose of *electrophoresis* is to determine the composition of the sample, i.e. detect the presence and relative amounts of the various alleles. This detection method enables a separation of the sample according to both the primer dye and fragment length, allowing the simultaneous detection of alleles across several loci, provided that different dyes are used for loci exhibiting alleles of similar length.

The idea behind electrophoresis is to let the molecules move through some separating medium, say a gel. The DNA is negatively charged once dissolved in a buffer and therefore moves from the negative to the positive electrode when an electric field is applied. The shorter molecules pass more easily through the gel, and the molecules are therefore separated according to length.

In capillary electrophoresis, the molecules move through a capillary filled with a gel-like substance. Laser light is shone onto a hole in the capillary, illuminating the dyes as they pass. A prism redirects the light to a detector picking up the signal from each dye separately. As the molecules pass the detector, the fluorescence intensity of the dyes – measured in *relative fluorescence units (RFU)* – is recorded together with the corresponding *data point* or *scan number* indicating the time since the start of the electrophoresis. In order to interpret the signal in terms of the presence of

alleles at STR loci, two steps are taken; the time variable determines the length of each molecule, and the length can in turn be used for determining the allele.

In an *electropherogram (EPG)*, such as Figure 1.8, we find fluorescence intensity plotted against fragment length in a separate panel for each dye. Using a user-specified choice amongst various smoothing methods, peaks in intensity are identified and the corresponding heights in RFU recorded. An important property of the EPG regarding interpretation of a DNA profile is that the *peak height* of an allele is roughly proportional to the amount of the allele in the sample; as is the *peak area* (see e.g. Tvedebrink et al. (2010)).

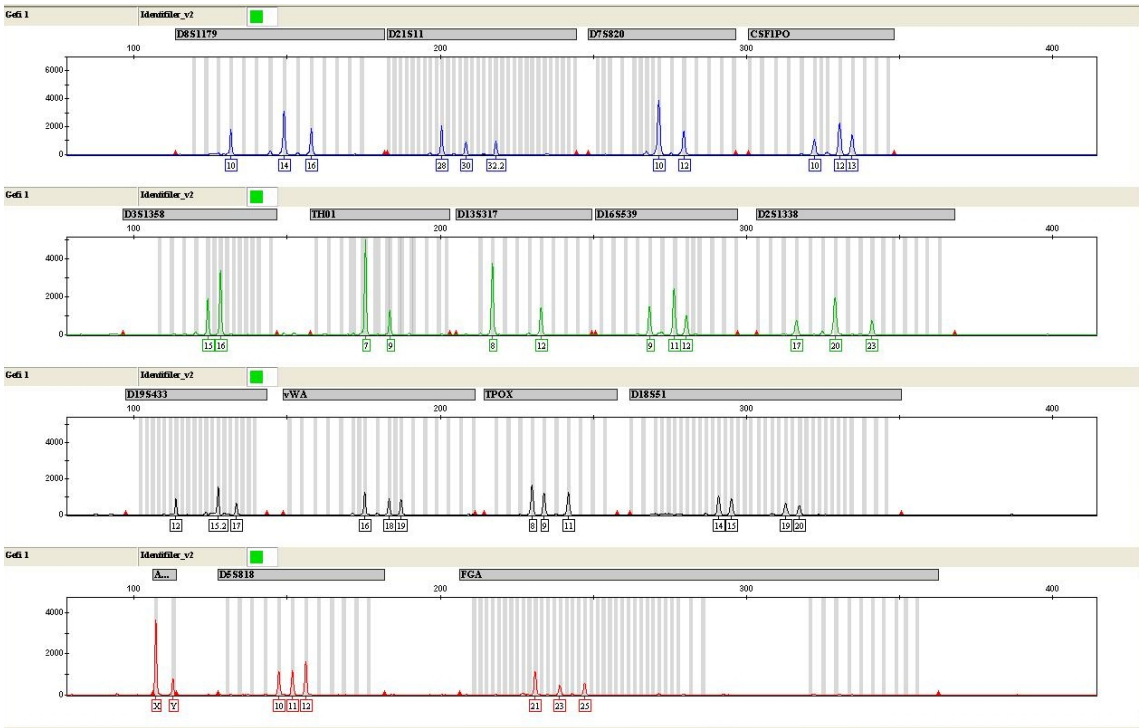


Figure 1.8: Electropherogram. Panels correspond to dyes. The horizontal axes indicate length in base pairs, and the vertical axes indicate peak height in RFU. Grey bars above each panel indicate the range of alleles corresponding to specific markers. The image is taken from Cowell et al. (2015).

Sizing of fragments

An *internal size standard* is a collection of DNA fragments of known lengths that are mixed into the DNA sample of interest before the amplification. To distinguish

the fragments in the size standard from the fragments in the sample of interest, a separate colour dye is used for their primers; when included in the EPG, the peaks for the size standard will be displayed in a separate panel.

By fitting a suitable curve, the relationship between the known fragment lengths and the observed data points for the internal size standard is established. This curve is then used to estimate the lengths of fragments in the sample of interest, by using the data points at which a peak has been observed.

Allele calling

Typically a batch of samples are analysed simultaneously in a tray with multiple wells. One of the wells contains a constructed sample, an *allelic ladder*, consisting of a collection of alleles of known lengths, which are then also amplified and sized.

The alleles in the ladder are then used to determine the alleles in the DNA sample, in that lengths for alleles in the sample are matched to bins created using the estimated lengths of the alleles in the ladder and a window size of typically ± 0.5bp. Any allele falling outside a bin is flagged as an *off-ladder* allele, and may then be determined manually. Not all possible alleles are included in the ladder, particularly microvariants may well not be included. Alleles that differ only a few base pairs in length from an allele in the ladder would typically be identified by interpolation.

Going back to the EPG in Figure 1.8 we see four panels, one for each of the four dyes used with the Identifiler™ kit: blue, green, yellow, and red; the panel corresponding to the orange dye used for the internal size standard is not displayed. The vertical axes indicate intensity in RFU, and the horizontal axes the length in base pairs. The allelic range for a locus is indicated at the top of each panel along with the locus name. The grey vertical bars correspond to bins for the alleles in the allelic ladder. Some of the peaks have been registered as alleles – these alleles have been *called* – and the repeat number is then displayed below the peak. The EPG is

typically summarised in terms of a set of variables describing the set of peaks; see Summary 1.1.

Summary 1.1: Peak data from an EPG

Dye	Colour of the dye of the fragment.
Data point	Time point at which the peak is detected.
Size	Fragment length measured in base pairs, as estimated from the data point.
Locus	Marker name, if the size falls within the range of an STR marker.
Allele	Allele name (repeat number), if the size corresponds to an allele.
Height	Peak height in RFU.
Area	Peak area.

In order to avoid recording peaks that are really background noise, usually some minimum threshold is set for the peak height. This is commonly referred to as a *peak amplitude threshold* or simply a *detection threshold*.

For a single-source DNA sample of high quality, generally the EPG is a unique representation of the person's DNA profile, and so the term DNA profile is naturally used interchangeably for the genetic markup and the EPG. In the case of multiple donors, we may use the term *mixed profile* to denote the EPG.

1.1.4 Artefacts

There are various reasons why sometimes the EPG does not give a true picture of the composition of the sample. We now describe some of the common *artefacts* that complicate the analysis of a DNA sample.

We shall focus on artefacts directly associated to missing or extra peaks in the EPG, thereby giving a false picture of the allele presence. However, there are other aspects that may be worth addressing in the future; for instance, if the DNA is degraded, the height of peaks is typically declining with the length of the allele.

Additional peaks in the EPG

During PCR, slippage of the two strands can result in misalignment of the template molecule and the copy, so that the copy becomes either a number of repeats too short or too long; see Figure 1.9.

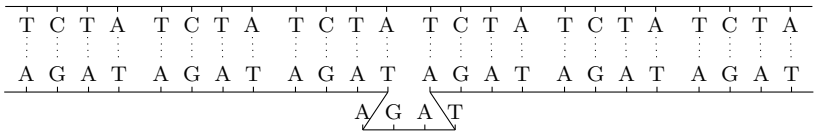


Figure 1.9: The strands are misaligned after reannealing.

The most common stutter product – and the only one we shall consider – is that of one repeat shorter than the template. Note that in the case of alleles with a complex repeat structure, the loss of one repeat motif for allele a may not result in an allele of the same DNA sequence as allele $a-1$. The length and dye colour of the stutter product, however, will be the same as that of an allele of type $a-1$, rendering it indistinguishable from actual alleles of type $a-1$. A *stutter peak* is a peak that is attributed to be solely due to the presence of stutter products. Often a heuristic filter is applied to remove stutter peaks, for instance a peak at $a-1$ may be classified as stutter if its height is less than, say, 15% of the peak at a.

Another possible reason for an extra peak is contamination with a few alleles, typically unrelated to the sample of interest, and this is called *dropin*.

Missing peaks in the EPG

When an allele that is present in the sample is for some reason is not detected through the presence of a peak, we say that it has *dropped out*. This happens, for instance, when

(i) the amount of DNA is too small for the peak to raise above the applied detection threshold.

(ii) the allele is not amplified (a *silent allele*), for instance due to a mutation in the primer binding site.

(iii) the allele size is outside the calling range for the particular locus and is not detected (a *null allele*).

1.2 DNA mixture analysis

We now turn to the problem of statistical analysis of DNA samples in the context of a crime case. We are particularly interested in the analysis of DNA mixtures, i.e. samples which contain DNA from multiple *contributors*. Commonly, a distinction is made in the literature between DNA mixtures and samples containing DNA from a single individual because different procedures are used for their analyses. However, for our purpose we may simply think of a single-source sample as a trivial mixture of DNA from a single contributor.

The analysis of a DNA mixture can have various objectives; common for these is determining the composition of the sample in terms of the DNA profiles of the contributors. We approach this problem by jointly modelling the DNA profiles of potential contributors to the sample and the peak heights in the EPG for the typed sample.

A model for a mixture naturally involves an explanation of the DNA sample in terms of a set of potential contributors. We distinguish between *known* and *unknown* contributors to the sample. The DNA profile of an unknown (unprofiled) contributor is given a distribution, typically mimicking the distribution of profiles in a suitable *reference population*. In contrast, the DNA profile of a known (profiled) contributor is considered fixed and known.

By *evidence* E we shall mean the set of peak heights and the DNA profiles of any profiled individuals associated with the case. Then, given the observed evidence E, we try to infer the composition of the mixture, which we may think of as the

set of DNA profiles and the proportions in which the DNA is mixed. In this thesis, the *mixture proportions* are considered fixed whereas the genotypes are random variables.

An explanation of the sample in terms of the contributors may be referred to as a *hypothesis* H and defines the joint model $\Pr(E, \boldsymbol{g} \,|\, H)$ of the evidence E and the full set of genotypes $\boldsymbol{g}$ constituting the set of DNA profiles for the contributors. This may be specified in two parts,

$$\begin{aligned} \Pr(E, \boldsymbol{g} \,|\, H) &= \Pr(E \,|\, \boldsymbol{g}, H) \Pr(\boldsymbol{g} \,|\, H) \\ &= \Pr(E \,|\, \boldsymbol{g}) \Pr(\boldsymbol{g} \,|\, H) \end{aligned} \tag{1.1}$$

Here one part is a specification of the distribution $\Pr(\boldsymbol{g} \,|\, H)$ of the DNA profiles of the unknown contributors, and the other part is a specification of the distribution $\Pr(E \,|\, \boldsymbol{g}, H)$ of peak heights given a specific allelic composition of the mixture. We make in (1.1) the standard assumption that $\Pr(E \,|\, \boldsymbol{g}, H) = \Pr(E \,|\, \boldsymbol{g})$, meaning that the distribution of peak heights only depends on the particular hypothesis H in question through the allelic composition of the sample. This is essentially an assumption about the mechanism generating the EPG; the EPG is a product of the molecular composition of the sample, and thus depends on each specific individual only through its genetic make-up, the DNA profile.

The marginal distribution of peak heights is then found by summing over all possible genotypes under the hypothesis H as

$$\Pr(E \,|\, H) = \sum_{\boldsymbol{g} \in \mathcal{G}} \Pr(E \,|\, \boldsymbol{g}) \Pr(\boldsymbol{g} \,|\, H). \tag{1.2}$$

1.2.1 Weight of evidence

In the context of a crime case there are typically two or more competing explanations of the DNA sample, and it is of interest to quantify the strength of evidence toward one explanation over the other.

A classical example is that of assessing the strength of the evidence against a specific person K having contributed to the sample. For this purpose we may formulate two hypotheses in the setting of a trial. The *prosecution* hypothesis H_p specifies that the DNA profile of the *defendant* K is present in the sample, i.e. K is among the known contributors. The competing explanation, the *defence* hypothesis H_d, typically replaces K with an unknown contributor U. Note that this alternative explanation does not exclude the possibility of U having the same DNA profile as K. The strength of the evidence against K is quantified in terms of a *likelihood ratio* (Good, 1950; Lindley, 1977; Balding, 2005),

$$LR = L(H_p)/L(H_d) = \Pr(E \mid H_p)/\Pr(E \mid H_d). \tag{1.3}$$

We follow Balding (2013) and Steele and Balding (2014) in reporting the *weight of evidence (WoE)* as

$$\mathrm{WoE}(K) = \log_{10} LR \tag{1.4}$$

in the unit *ban* as introduced by Alan Turing (Good, 1979). A change of 1 ban in WoE thus represents a factor 10 change on the likelihood ratio.

1.2.2 Mixture deconvolution

The purpose of *mixture deconvolution* is to identify the DNA profiles of the contributors to the stain. This includes determining the number of contributors and their genotypes, and it may also be of interest to determine the proportion of DNA from each contributor.

There are various approaches to deconvolution of a crime scene profile. One approach, which we adopt here, is to take the consequence of assuming a distribution for the DNA profiles and consider the posterior distribution $\Pr(\boldsymbol{g} \mid E, H)$ of configurations $\boldsymbol{g}$ of DNA profiles given the observed evidence E and the hypothesis H under consideration.

From a statistical point of view, the distinction between deconvolution and assessing the weight of evidence toward a specific person is perhaps not clear cut. For instance, it may well be of interest to investigate how the DNA profiles are explained under a defence hypothesis involving unknown contributors.

A natural prediction for the profile of an unknown contributor U would be a profile K maximising $\Pr(U = K \mid E, H)$. Here we use the common shorthand notation U for the random DNA profile of an unknown contributor under H and K for a fixed and known profile.

1.2.3 Evidential efficiency

In the special case of analysing the EPG from a high-quality single-source stain where the DNA profile of the contributor K is uniquely determined, the only DNA profile assigning the evidence a non-zero likelihood would be the profile of K. Consequently, $\Pr(E \mid H_p) = \Pr(E \mid K) = 1$, and

$$\begin{aligned}\Pr(E \mid H_d) &= \sum_{K'} \Pr(E \mid K') \, \mathrm{P}(U = K' \mid H_d) \\ &= \Pr(E \mid K) \, \mathrm{P}(U = K \mid H_d) \\ &= \mathrm{P}(U = K).\end{aligned}$$

The likelihood ratio (1.3) thus reduces to the inverse *match probability*, i.e. the probability $\mathrm{P}(U = K)$ that a random individual has the same profile as K.

Consider again the evidence against a person K based on comparing a hypothesis H_p, in which K is included, to the hypothesis H_d formed by replacing K by an unknown contributor U; here, H_p can simply be formulated as H_d where it is known that $U = K$.

We can then conclude that basing the evidence on a mixed sample can – regardless of the specific model applied – not lead to stronger evidence than that based on

high-quality sample, where K is the only donor:

$$\begin{aligned} LR &= \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)} \\ &= \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d, U = K)\,\mathrm{P}(U = K) + \mathrm{P}(U \neq K)\,\Pr(E \mid H_d, U \neq K)} \\ &= \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_p)\,\mathrm{P}(U = K) + \mathrm{P}(U \neq K)\,\Pr(E \mid H_d, U \neq K)} \\ &\leq \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_p)\,\mathrm{P}(U = K)} = \frac{1}{\mathrm{P}(U = K)}. \end{aligned} \tag{1.5}$$

In particular, we have that

$$\mathrm{WoE}(K) = \log_{10} LR \leq -\log_{10} \mathrm{P}(U = K). \tag{1.6}$$

This establishes an upper bound $\mathrm{WoE}_{\max}(K) = -\log_{10} \mathrm{P}(U = K)$ for the weight of evidence against K.

The 10 STR markers used for the SGM Plus® system typically lead to match probabilities in the order of magnitude of 10^{-14} (Applied Biosystems, 2012b), and so the WoE based on a high-quality single-source sample will be in the order of 14 bans; the newer system NGM™ uses 15 STR markers, which strengthens the evidence to the order of 20 bans (Applied Biosystems, 2012a). The weaker evidence obtained from a mixture in comparison to a high-quality sample can be seen as effectively reducing the number of markers used for the identification, as discussed in Cowell et al. (2015).

We can use the upper bound for the WoE to measure the quality of a particular DNA sample by defining the *loss of evidential efficiency (WL)* against K:

$$\mathrm{WL}(E \mid K) = \mathrm{WoE}(K)_{\max} - \mathrm{WoE}(K). \tag{1.7}$$

The loss is non-negative and a high-quality single-source sample will lead to a loss of 0. Note that both $\mathrm{WL}(E \mid K)$ and $\mathrm{WoE}(K)_{\max}$ are measured relative to the particular defence hypothesis under consideration.

Relation between WoE and deconvolution

Let us consider the problem of identifying an unknown contributor U under some hypothesis H and imagine that we have a suggested profile K. In assessing the suggested profile, we may evaluate the weight of evidence against an individual with profile K by comparing the hypothesis $H_d = H$ to the hypothesis H_p replacing U by the profile K. The loss of evidential efficiency against K is then

$$\begin{aligned}
\mathrm{WL}(E \mid K) &= \mathrm{WoE}(K)_{\max} - \mathrm{WoE}(K) \\
&= -\log_{10} P(U = K) - \log_{10} \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)} \\
&= -\log_{10} \frac{\Pr(E \mid H_p) P(U = K)}{\Pr(E \mid H_d)} \\
&= -\log_{10} \frac{\Pr(E \mid U = K, H_d) P(U = K \mid H_d)}{\Pr(E \mid H_d)} \\
&= -\log_{10} \Pr(U = K \mid E, H_d). \qquad (1.8)
\end{aligned}$$

Thus, the loss of evidential efficiency against a person with profile K can be expressed in terms of the posterior probability of U sharing the profile K.

The deconvolution of the mixture uniquely determines the genotypes of the individual exactly when there is no evidential loss. The smallest evidential loss – or strongest possible evidence – is obtained when K is the posterior most likely profile given the evidence, i.e. maximising $\Pr(U = K \mid E, H_d)$. The quantity $\mathrm{WL}(E \mid K)$ can therefore be seen as a measure of the *generic loss of evidential efficiency*.

Maximised likelihoods

When there are additional parameters in the model, the likelihood is of the form $\Pr(E \mid \psi, H)$. One approach to handling the unknown ψ is to estimate by the method of maximum likelihood.

Using maximised likelihood ratios preserves the property that the WoE against an individual K in a mixed trace can at most be that obtained by a perfectly matching single-source DNA profile, irrespective of whether the same estimates are

used for both hypotheses. To see this, let as before H_p be the prosecution hypothesis, in which K is included as a contributor and let H_d be the defence hypothesis that replaces K by a random individual U. Let further $\hat{\psi}_p$ be the maximum likelihood estimates of the unknown parameters ψ under the prosecution hypothesis and $\hat{\psi}_d$ the estimates under the defence hypothesis. We then have

$$\begin{aligned}\mathrm{WoE}(K \mid \hat{\psi}_d, \hat{\psi}_d) &= \log_{10} \frac{\Pr(E \mid H_p, \hat{\psi}_d)}{\Pr(E \mid H_d, \hat{\psi}_d)} \leq \log_{10} \frac{\Pr(E \mid H_p, \hat{\psi}_p)}{\Pr(E \mid H_d, \hat{\psi}_d)} \\ &= \mathrm{WoE}(K \mid \hat{\psi}_p, \hat{\psi}_d) \leq \log_{10} \frac{\Pr(E \mid H_p, \hat{\psi}_p)}{\Pr(E \mid H_d, \hat{\psi}_p)} \\ &= \mathrm{WoE}(K \mid \hat{\psi}_p, \hat{\psi}_p) \leq -\log_{10} \mathrm{P}(U = K) = \mathrm{WoE}(K)_{\max},\end{aligned}$$

where the last inequality follows from (1.6).

In this thesis, we estimate under each hypothesis separately, allowing both sides to obtain as favourably a likelihood as possible. Using solely the estimates obtained under the defence hypothesis weakens the evidence, and thus favours the defence. Note that it may be of interest to consider different types of defence hypotheses than those replacing the defendant by an unknown contributor; for instance, the number of contributors could be different under the prosecution and the defence hypotheses. In such cases, it may not be meaningful to use the same set of parameters under both hypotheses.

Several unknown contributors

In the arguments above, we have considered a defence hypothesis that replaces K with one particular unknown contributor U. However, often the defence hypothesis involves several unknown contributors and it is then unclear which of these should be considered the substitute for K.

Assuming that the defence hypothesis involves N unknown contributors, whose DNA profiles are independent and identically distributed, H_p may be phrased instead as H_d with *one or more* of the N unknown contributors sharing the profile of K. The

argument may then be carried out as in (1.5), simply by replacing the probability of a match by the – larger – probability $1 - (1 - \mathrm{P}(U = K))^N$ of a match to one or more of the unknown contributors. We therefore obtain a smaller maximal WoE,

$$\mathrm{WoE}(K) \leq -\log_{10}(1 - (1 - \mathrm{P}(U = K))^N) \leq -\log_{10} \mathrm{P}(U = K),$$

but note that this implies that the upper bound (1.6) still holds.

1.3 An example case: MC15 and MC18

We shall use as a running example a case discussed in Gill et al. (2008). The case involves two DNA samples *MC15* and *MC18*, which are both believed to contain DNA from at least three contributors. Related to the case we have the DNA profile of the victim K_1 and also those of a further two genotyped individuals K_2 and K_3. The full peak height data and DNA profiles of the three individuals can be found in Table A.1.

The case is analysed in both Cowell et al. (2015) and Graversen and Lauritzen (2014). Further the case is used as an example throughout the documentation for DNAmixtures (Graversen, 2014).

Figures 1.10, 1.11, and 1.12 show the observed peak heights for the two samples in terms of stylized EPGs corresponding to the blue, green, and yellow dye panels. We have for each marker let the horizontal axes correspond to the allele repeat number rather than allele length in base pairs. Microvariants with incomplete repeats .1, .2, or .3 are marked using quarter lengths of whole repeats.

The two samples are similar both in terms of which alleles are observed and in terms of the proportion of each allele, as represented by the relative height of the peak. The horizontal lines mark the threshold of 50 RFU, which we shall use throughout our analyses, as this is a commonly applied threshold and has already been applied to the peak heights of MC15 in Gill et al. (2008). Two peaks for MC18,

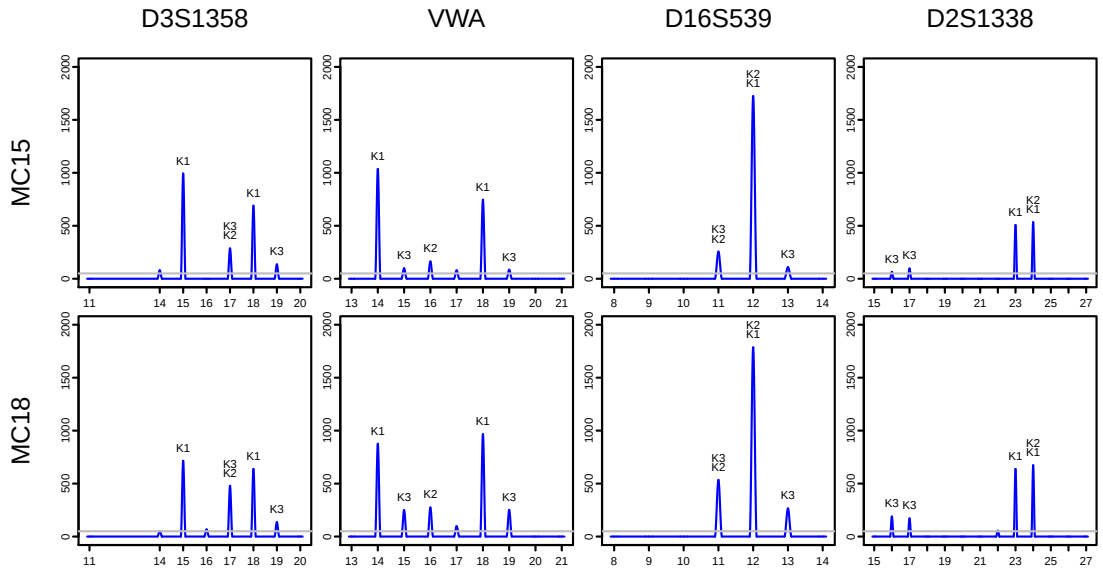


Figure 1.10: Peak heights for markers in the blue dye lane, as obtained for the samples MC15 and MC18. The horizontal lines mark a detection threshold of 50 RFU.

marker FGA (Figure 1.12) fall below the threshold – allele 21 with a peak height of 49 RFU, and allele 25 with a peak height of 39 RFU.

The hypotheses of interest revolve around explaining the sample by the presence of one or more of K_1, K_2, and K_3. All three dyes exhibit some peaks that are not explained by any of the three individuals, for example allele 14 for marker D3S1358 (Figure 1.10). All of these peaks can be explained by stutter from an allele among the three individuals; for instance, allele 14 would be stutter from the allele 15, which is an allelic peak as it is in the genotype of K_1. Assuming that K_2 is indeed present in the sample MC15, a few alleles for the markers with yellow dye (Figure 1.12) have dropped out: allele 16.2 for D19S433, allele 9 for TH01, and allele 22 for FGA.

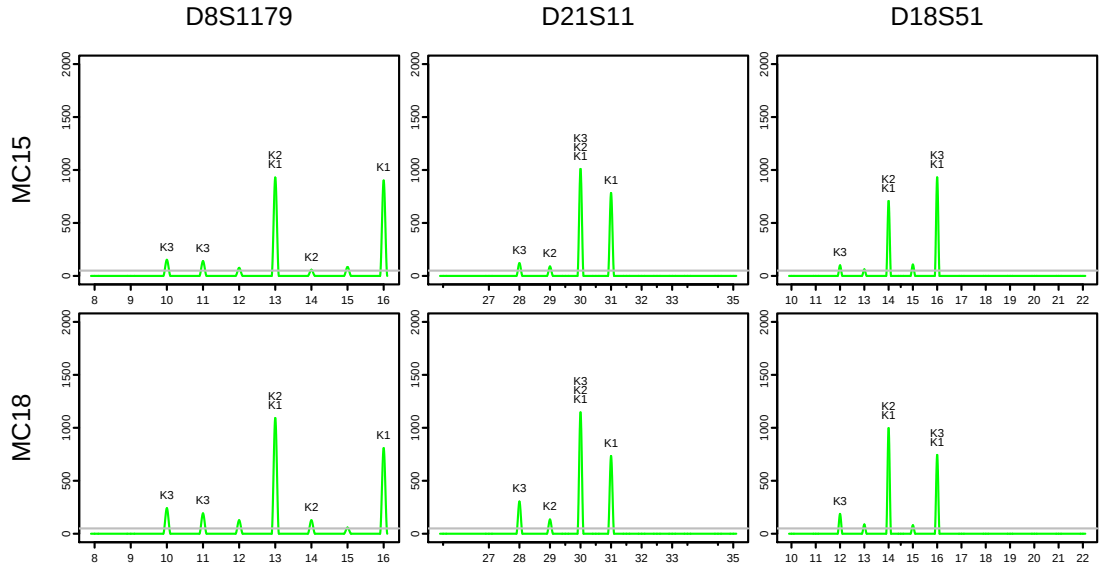


Figure 1.11: Peak heights for markers in the green dye lane. The horizontal lines mark a detection threshold of 50 RFU.

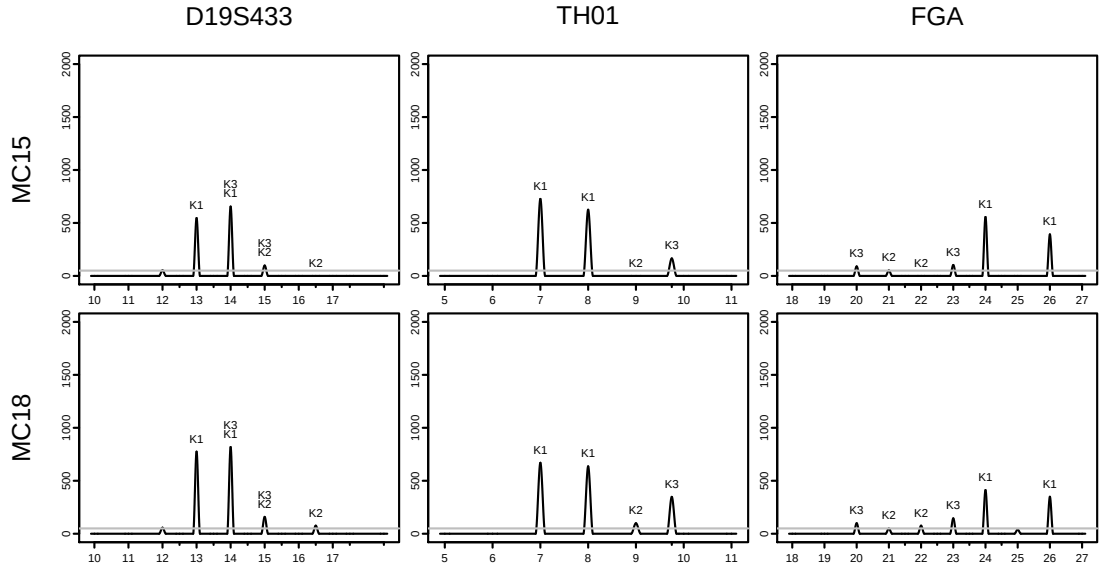


Figure 1.12: Peak heights for markers in the yellow dye lane. The horizontal lines mark a detection threshold of 50 RFU.

Chapter 2

A statistical model for DNA mixtures

This chapter sees the development of a model for DNA mixtures in terms of a joint model for DNA profiles of the contributors and the peak heights of the EPG.

We assume that a suitable range of possible alleles has already been established for each marker in the EPG, and that the point of interest lies in modelling the presence or absence of these alleles in the profiles of the contributors as well as the height of the peaks at these positions in the EPG. We denote by $a = 1, \ldots, A_m$ the range of alleles at marker m.

Following (1.1), the model $\Pr(E, \boldsymbol{g} \mid H)$ is composed of two parts. The first part, $\Pr(\boldsymbol{g} \mid H)$, is concerned with modelling the DNA profiles of contributors to the mixture (Section 2.1). The second part $\Pr(E \mid \boldsymbol{g})$ – our main focus – models the peak height variability for a mixture of known composition (Section 2.2).

2.1 Statistical model for DNA profiles

For a given hypothesis H, we need to specify a model $\Pr(\boldsymbol{g} \mid H)$ for the set of genotypes

$$\boldsymbol{g} = \{g_{im}; i = 1, \ldots, k; m = 1, \ldots, M\} \tag{2.1}$$

constituting the full set of DNA profiles based on M markers for all contributors i under H.

We adopt the standard model for DNA profiles, under which the unknown contributors are thought of as unrelated – also to known contributors associated to the case – and with DNA profiles distributed according to a suitably chosen reference population.

Formally, the standard model assumes mutual independence of the genotypes g_{im} in $\boldsymbol{g}$. This implies mutual independence of genotypes both within and between DNA profiles for the contributors. The standard model further considers the DNA profiles of known contributors to be fixed, whereas the two alleles of an unknown person are considered sampled independently from the reference population, i.e. the population is assumed to be in Hardy–Weinberg equilibrium.

2.1.1 Allele counts and allele frequencies

Consider a marker with a range $a = 1, \ldots, A_m$ of possible alleles. Letting n_{ia}^m denote the number of alleles a that contributor i possesses, the pair of alleles constituting the genotype at that marker can be represented by the vector of *allele counts* $(n_{i1}^m, \ldots, n_{iA_m}^m)$. We shall use the terms "genotype", "pair of alleles", and "allele counts" interchangeably since there is a one-to-one relation between any of two of these representations.

Using the allele-count representation, the full set of genotypes $\boldsymbol{g}$ for all individuals included under H is expressed as

$$\boldsymbol{n} = \{n_{ia}^m, i = 1, \ldots, k; m = 1, \ldots, M; a = 1, \ldots, A_m\}. \tag{2.2}$$

A consequence of the standard model is that a genotype $(n_{i1}^m, \ldots, n_{iA_m}^m)$ for an unknown contributor follows a multinomial distribution with $\sum_a n_{ia}^m = 2$ and *allele frequencies* $(q_1^m, \ldots, q_{A_m}^m)$. It is customary to assume population allele frequencies to be known and equal to values obtained from a database of individuals from the

reference population. One common choice of estimates is the set of empirical frequencies of alleles in the reference database, which we shall use throughout unless otherwise stated.

2.1.2 F_{ST} and sampling-adjustment parameters

One particularly common variation to using the empirical allele frequencies is the use of *adjusted* allele frequencies (Balding, 2013), which in effect corresponds to using empirical frequencies after adding the profile of the defendant to the database, possibly with a weight F_{ST} as a way of taking distant relatedness into account.

Estimated frequencies are often based on rather few individuals (~ 300), leading to some alleles having an estimated frequency of zero or close to zero; commonly, pseudo counts are added to the database to bias up the empirical frequencies for such alleles. This *sampling adjustment* may include adding the entire crime scene profile, adding only the profile of the defendant, or adding all profiled individuals.

Any of the mentioned adjustments can readily be incorporated in our framework as it is simply a matter of using a different set of estimates for the allele frequencies in the standard model.

It should be noted that many of this type of adjustments depend on the specific hypotheses and designated defendant(s) under consideration and that the choice of allele frequencies and their adjustment is important for real casework. For simplicity and uniformity we use the raw allele frequencies throughout the example analyses in Chapters 5-6 unless otherwise stated.

2.2 Peak-height distribution for DNA samples of known composition

In the following, we develop a model $\Pr(E \mid \boldsymbol{g})$ for the peak heights across all markers and alleles in the EPG for a mixture of known composition; in particular, the

DNA profiles $\boldsymbol{g}$ and the proportions in which the contributors' DNA occur are here considered fixed.

The model is an extension of the basic gamma model presented in Cowell et al. (2007a) and revisited in Section 2.2.1 below. The model for peak heights is then elaborated in Section 2.2.2 to allow for the presence of stutter, and finally in Section 2.2.3 we introduce dropout as the event of a peak falling below a chosen minimum threshold.

The gamma model was also extended in Cowell et al. (2011), but in a direction different from here; their model has a more complicated stutter component, and dropout is modelled solely as a pre-PCR sampling phenomenon. The differences are briefly discussed in Section 9.1.2, where also another extension of the gamma model as found in Puch-Solis et al. (2013) is briefly touched upon. An advantage of our model is that it allows some computational simplifications, making it feasible to handle more complex scenarios such as those involving a higher number of unknown potential contributors.

2.2.1 The gamma model

We adopt the basic gamma model of Cowell et al. (2007a) in the use of gamma distributions to capture the variability in peak heights; however, we deviate by directly modelling the absolute peak heights. Thus, we do not normalise the peak heights, nor do we take into account the effect of the length of the allele on the resulting peak heights.

Motivation for gamma parameters

An important property of peak heights (and areas) is that a peak height is approximately proportional to the quantity of the allele in the mixture (see e.g. Tvedebrink et al. (2010)), which we may express as a proportionality between mean peak height and quantity. Furthermore, Tvedebrink et al. (2010) indicated that the variance

of the peak height is proportional to the mean peak height with a proportionality factor that does not depend on the quantity; this was also an underlying assumption of the gamma model of Cowell et al. (2007a).

In our development of the model below, we shall for the sake of reasoning take the quantity of a particular allele to be directly proportional to the number of molecules.

Now, assuming that the amplification of γ molecules of the same type leads to a (random) peak height of $H(\gamma)$, the proportionality properties are naturally expressed as $\mathbb{E}\, H(\gamma) = c\gamma$ and $\mathbb{V}\, H(\gamma) = \eta c\gamma$, where c and η are constants that do not depend on γ. Assuming that $H(\gamma)$ follows a gamma distribution, then in order to respect the proportionality we must for any γ have that

$$H(\gamma) \sim \Gamma(\gamma c/\eta, \eta). \tag{2.3}$$

where $\Gamma(\alpha, \beta)$ denotes the distribution with density

$$g(h \,|\, \alpha, \beta) = \frac{h^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-h/\beta} \text{ for } h > 0.$$

We let $\Gamma(0, \beta)$ denote the distribution degenerate at 0, covering the case where there is no DNA and therefore no peak with probability 1. We see that the shape parameter should be proportional to the amount γ, whereas the scale parameter η should be the same for all amounts γ.

Modelling a specific mixture

Consider now a DNA mixture of k potential contributors, where each individual i has contributed some fixed fraction $\phi_i \geq 0$ of DNA. The vector $\phi = (\phi_1, \phi_2, \ldots, \phi_k)$ then denotes the fractions from all contributors prior to amplification and respects that $\sum_{i=1}^{k} \phi_i = 1$. It is assumed that these fractions of DNA are constant across markers; this corresponds to seeing ϕ_i as the fraction of whole cells from individual i in the mixture, and all further variability in the peak heights arising from various sampling procedures is then represented by the variability of the gamma distribution.

It is important that mixture proportions are constant across markers, as the unknown contributors are essentially identified by their contribution of DNA. In contrast, one would typically expect other parameters relating to the amplification process to be marker dependent and possibly dependent on fragment length in case of degraded DNA (Tvedebrink et al., 2012c; Balding, 2013; Puch-Solis et al., 2013).

Considering a marker m and letting γ be a measure of the total number of cells in the mixture prior to amplification, there are $\gamma \sum_{i=1}^{k} \phi_i n_{ia}^m$ alleles of type a in the mixture prior to amplification: A single individual i has contributed with a fraction ϕ_i of the γ cells, and for each of these the individual has $n_{ia}^m \in \{0, 1, 2\}$ alleles of type a; the individual has therefore contributed $\gamma \phi_i n_{ia}^m$ alleles of type a and the total number of alleles a in the mixture is thus $\gamma \sum_{i=1}^{k} \phi_i n_{ia}^m$.

Introducing the *effective number of alleles of type a*,

$$B_a^m(\phi, \mathbf{n}) = \sum_i \phi_i n_{ia}^m, \tag{2.4}$$

we can think of the molecules in the mixture as originating from a single contributor with $B_a^m(\phi, \mathbf{n})$ alleles of type a and a total of $\sum_a B_a^m(\phi, \mathbf{n}) = 2$ alleles per marker, since $\sum_i \phi_i = 1$. If ϕ is interpreted as a probability distribution, $B_a^m(\phi, \mathbf{n})$ is the expected number of alleles of type a per cell.

From (2.3) it is clear that if we wish to capture variability in the peak heights by gamma distributions, then the shape parameter for the peak height for allele a should be $\gamma B_a^m(\phi, \mathbf{n}) c/\eta$ for a suitable proportionality factor c. For convenience, we define the parameter $\rho = \gamma c/\eta$, which is then proportional to the number γ of cells in the mixture.

Model 2.1: The basic gamma model

Consider a crime scene profile covering markers $m = 1, \ldots, M$, each of which has an allelic range $a = 1, \ldots, A_m$. The model describes the variability in the set of observed peak heights

$$\{H_a^m; m = 1, \ldots, M; a = 1, \ldots, A_m\}.$$

Given the full set of genotypes $\boldsymbol{n}$ for all individuals, the peak heights are mutually independent and gamma distributed as

$$H_a^m \,|\, \boldsymbol{n} \sim \Gamma\left(\rho \sum_{i=1}^{k} \phi_i n_{ia}^m,\ \eta\right). \tag{2.5}$$

In a single-source sample with no artefacts and a heterozygous contributor, where then $n_{1a}^m = n_{1b}^m = 1$ for two alleles $a \neq b$ at marker m, we have that $\mu = \rho\eta$ is the mean peak height. Further, in this case $\sigma = 1/\sqrt{\rho}$ is the coefficient of variation for peak heights at that marker and may thus be interpreted as a measure of generic peak imbalance in that a high σ corresponds to a high variability in peak heights and thus a higher probability of more extreme imbalance. In the presentation of results we shall often use (σ, μ) instead of (ρ, η) because of their more direct interpretability.

For notational simplicity, in the following we suppress the dependence of various quantities on markers m unless it would lead to confusion.

Decomposition of peak heights

It should be noted that in contrast to Cowell et al. (2015), we do not decompose peak heights into a sum of independent individual contributions H_{ia} from each contributor i; we are only concerned with the distribution of the total peak heights, and we do not need assumptions about these unobservable individual contributions.

A consequence of using the gamma distribution for modelling peak heights is that it enables a decomposition into independent gamma-distributed contributions.

In principle we could use other families of distributions in place of the gamma distribution, and for this it is important that our methodology does not rely on an assumption about the individual contributions.

Heterozygous balance

Consider a single-source sample and a marker, at which the donor is heterozygous. The two alleles at this marker are present in the same amount and are therefore expected to exhibit peaks of similar size. A commonly used quantity in the interpretation of DNA profiles is the *heterozygote balance* $\text{Hb} = H_a/H_{a'}$ between two peaks for a single, heterozygous donor with genotype (a, a').

There are various definitions of the heterozygote balance regarding the order of the two alleles when taking the ratio; we shall define here Hb to be the ratio of the peak height for the longer allele a to that of the shorter allele a', and so $\text{Hb} \in (0, \infty)$ with $\text{Hb} = 1$ corresponding to peaks of identical height.

Bounds on the heterozygote balance are used as a heuristic, either directly in the interpretation of the profile, or as an ad-hoc method for limiting the number of possible allocations of genotypes to unknown contributors (Tvedebrink et al., 2010; Puch-Solis et al., 2013). One commonly used range is that of Bill et al. (2005), giving a guideline of $0.6 \leq \text{Hb} \leq 1/0.6$ for acceptable values of the heterozygote balance.

A consequence of the gamma model for peak heights (ignoring stutter and dropout) is that Hb follows an F-distribution with both parameters equal to $2\rho = 2/\sigma^2$. As illustrated in Figure 2.1, the imbalance increases with the increased variability in peak heights, particularly as the amount of DNA gets smaller. The mean heterozygote balance is for $\rho > 1$ $(\sigma < 1)$

$$\mathbb{E}\,\text{Hb} = \frac{\rho}{\rho - 1} = \frac{1}{1 - \sigma^2} \geq 1,$$

and the increasing trend for larger peak variation is seen in Figure 2.1. When $\rho < 1$ ($\sigma > 1$), large values of Hb become increasingly prominent and $\mathbb{E}\,\mathrm{Hb}$ becomes infinite. The studies of Bright et al. (2011, 2012) show that peak imbalance is more

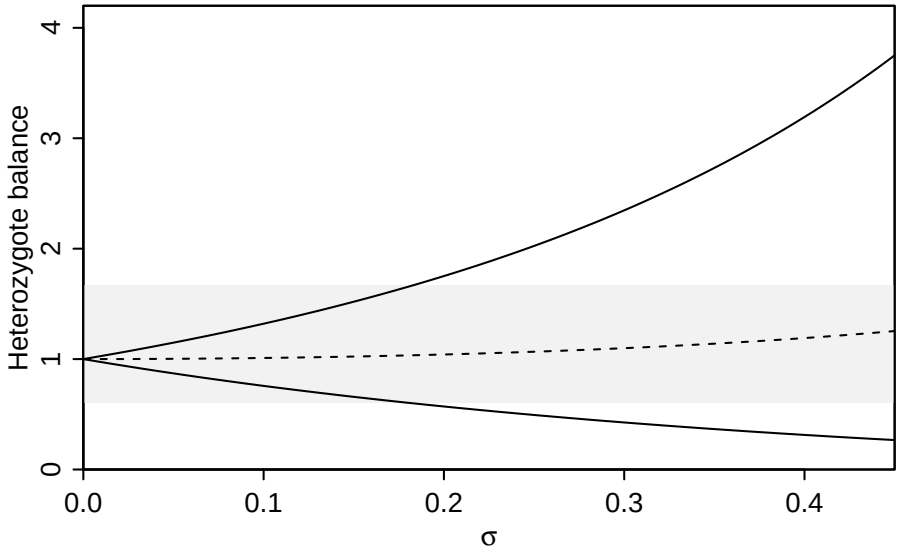


Figure 2.1: Pointwise 95% prediction bands (solid) for the heterozygote balance Hb and the expected Hb (dashed) in relation to $\sigma = 1/\sqrt{\rho}$. The horizontal band indicates the range $(0.6, 1/0.6)$.

exacerbated when the average (observed) peak height or area is small. For our model, the theoretical mean peak height is $\mu = \rho\eta$, and as Hb decreases with ρ and is constant in η the model respects that Hb decreases with an increasing mean peak height; the relationship is seen in Figure 2.2. The funnel shape characterising the ranges of Hb is also consistent with the study of Bright et al. (2011, 2012).

Bright et al. (2011) suggest that the guideline $\mathrm{Hb} \in (0.6, 1/0.6)$ of Bill et al. (2005) only applies to average peak heights above 2000 RFU, and that for average peak heights in the range 400-2000 RFU a range of $\mathrm{Hb} \in (0.5, 2)$ is more appropriate. Bright et al. (2012) more conservatively suggest that the range $\mathrm{Hb} \in (0.6, 1/0.6)$ is appropriate for samples with a mean peak height above 300 RFU.

Evidently such studies are dependent on the laboratory procedures and also our mean peak height μ is not directly comparable to the average peak height, which is an observed quantity. Bearing these limitations in mind, our model seems to

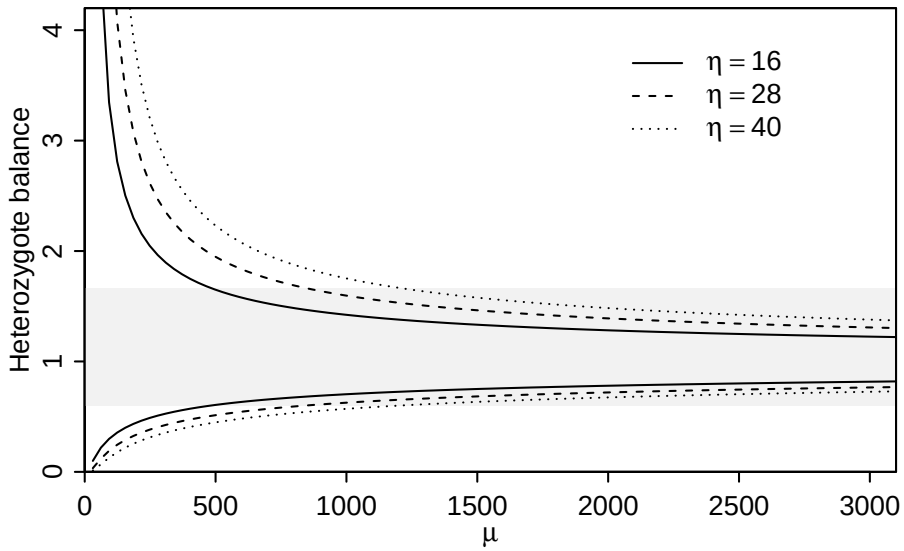


Figure 2.2: Pointwise 95% prediction bands for the heterozygote balance as a function of $\mu = \rho\eta$. The horizontal band indicates the range $(0.6, 1/0.6)$.

conform well with the behaviour discussed in Bright et al. (2011, 2012) and Bill et al. (2005).

2.2.2 Stutter

As described in Section 1.1.4, amplification of alleles of type a will in practice give rise to some stutter products that are one repeat motif shorter than the template alleles. Since such stutter products are of the same length as alleles of type $a-1$ and also labelled with the same colour dye, they will contribute to the peak at $a-1$ along with any existing alleles of type $a-1$. Thus, the peak for allele a may be inflated by stutter from any alleles $a+1$ and deflated by stutter to allele $a-1$. We ignore the possibility of stutter from allele a affecting any other peaks than that of $a-1$.

The gamma distribution (2.5) for peak heights is modified here to capture the "rearranging" of signal in the EPG. We do this through modification of the effective number of alleles, (2.4), resulting in the *effective number of alleles after stutter*

$$D_a^m(\phi, \xi, \mathbf{n}) = (1-\xi)B_a^m(\phi, \mathbf{n}) + \xi B_{a+1}^m(\phi, \mathbf{n}). \tag{2.6}$$

where $\xi \in [0, 1)$ is the *mean proportion of stutter*. We may think of ξ as the effective proportion of molecules that leads to stutter products during amplification; "effective", as evidently a specific molecule will not give rise to only stutter products.

Under the modified gamma model, peak heights are – conditionally on genotypes $\boldsymbol{n}$ – mutually independent and gamma distributed as

$$H_a^m \mid \boldsymbol{n} \sim \Gamma\{\rho D_a^m(\phi, \xi, \mathbf{n}), \eta\}. \tag{2.7}$$

Note that the basic gamma model (Model 2.1) simply corresponds to the special case $\xi = 0$.

Interpretation in terms of decomposition of peak heights

Assuming a common scale parameter η for peak heights H_a for different alleles a within a marker enables a simple interpretation of the model in terms of adding independent contributions to peaks both from proper alleles and from stutter. Imagine that the total signal from alleles of type a is distributed as in (2.5), but that this signal will not all appear as a peak for allele a; rather, it is decomposed into two independent contributions H_a^s and H_a^0, where H_a^s represents the part of the EPG signal originating from stutter products, and H_a^0 represents the remainder that goes through the PCR process undamaged. Letting the two components be independent and gamma distributed as

$$H_a^s \sim \Gamma\{\rho\xi B_a(\phi, \boldsymbol{n}), \eta\}, \quad H_a^0 \sim \Gamma\{\rho(1-\xi) B_a(\phi, \boldsymbol{n}), \eta\},$$

the total peak height $H_a = H_a^0 + H_{a+1}^s$ observed at allele a is gamma distributed as in (2.7). A peak in stutter position may itself include contributions from proper alleles (the stutter masks alleles from donors), and the parent peak may itself include stutter contributions from an allele with higher repeat number. As the contributions are mutually independent and gamma distributed with the same scale parameter,

the relative contribution lost to stutter

$$X_a = \frac{H_a^s}{H_a^s + H_a^0}$$

follows a beta distribution $\mathcal{B}\{\xi\rho B_a(\phi, \mathbf{n}), (1-\xi)\rho B_a(\phi, \mathbf{n})\}$ with mean $\mathbb{E}(X_a) = \xi$.

In the simple event that a peak is entirely due to stutter from a parent peak, which has not itself been inflated by stutter, we have $H_a = H_a^0$ and $H_a^s = H_{a-1}$ and can thus observe the effect of stutter directly. In this case, $X_a = H_{a-1}/(H_{a-1} + H_a)$ is the observed ratio of the stutter peak to the total signal and ξ is the expected proportion of the signal appearing as stutter. Note that we consider proportion of stutter out of the total height of the stutter and the main peak, whereas standard practice measures the stutter peak in percent of the main peak (Butler, 2005).

Our peak height model evidently also implies a model for the standard definition $H_{a-1}/H_a = X_a/(1 - X_a)$ of observed stutter, and we may for instance use this to compute prediction intervals for the amount of stutter. For a model with just one heterozygote donor and parameters $\sigma = 0.18$ and $\xi = 0.07$, a 95% prediction interval indicates that a stutter peak would be 1-22% the height of the parent peak. We have here used values for σ and μ that are representative of the analyses in Chapter 5 based on MC15 and MC18. The implied model for stutter respects that the variability in the amount of stutter is larger for small amounts of DNA, i.e. large $\rho B_a(\phi, \mathbf{n})$.

It should be noted that a consequence of the mutual independence between peak heights is that – given genotypes and model parameters – the height H_{a-1} of a pure stutter peak is independent of the height H_a of the parent peak. This conditional independence assumption is made also by Puch-Solis et al. (2013) and Taylor et al. (2013), the latter noting that it is supported by their empirical data.

Interpretation in terms of a beta distributed fraction lost to stutter

We may also think of the model directly in terms of the basic gamma model (2.5), where peak heights H_a are rearranged by a set of independent beta-distributed variables $X_a \sim \mathcal{B}\{\xi\rho B_a(\phi, \mathbf{n}), (1-\xi)\rho B_a(\phi, \mathbf{n})\}$, so that the observed peak at allele a is $(1-X_a)H_a + X_{a+1}H_{a+1}$ rather than H_a.

This interpretation of the model relies on the sum of the parameters in the beta distribution of X_a being equal to the shape parameter of $H_a \sim \Gamma(\xi\rho B_a(\phi, \mathbf{n}), \eta_a)$, as only then may the gamma-distributed peak height be decomposed into two independent gamma-distributed parts

$$X_a H_a \sim \Gamma(\xi\rho B_a(\phi, \mathbf{n}), \eta_a)$$
$$(1-X_a)H_a \sim \Gamma\{(1-\xi)\rho B_a(\phi, \mathbf{n}), \eta_a\}.$$

Secondly, if the scale parameter is the same across alleles, the sum $(1-X_a)H_a + X_{a+1}H_{a+1}$ is gamma distributed with the distribution (2.7).

2.2.3 Dropout

We allow a detection threshold $C \geq 0$ to be applied, in which case the peak height H_a^m is only observable through the censored peak height

$$Z_a^m = \begin{cases} H_a^m, & H_a^m \geq C \\ 0, & H_a^m < C. \end{cases}$$

We shall simply speak of Z_a^m as the peak height for allele a, and this may then be 0 either because there are no alleles a in the DNA mixture, or because no peak is observed above the threshold.

Denoting by $\boldsymbol{n}_a^m = (n_{ia}^m, i = 1, \ldots, k)$ the vector of allele counts for allele a across individuals, the distribution of Z_a^m given genotypes $\boldsymbol{n}$ depends only on the genotypes

through $\boldsymbol{n}_a^m$ and $\boldsymbol{n}_{a+1}^m$. For $a = 1, \ldots, A_m$ the distribution is a mixture with density

$$f_\psi(z_a^m \,|\, \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m) = \begin{cases} g_\psi\Big(z_a \,\Big|\, \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m\Big), & z_a^m \geq C \\ G_\psi\Big(C \,\Big|\, \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m\Big), & z_a^m = 0 \end{cases} \tag{2.8}$$

where g denotes the density and G the cumulative distribution function for the gamma distribution as in (2.7), and $\psi = (\phi, \xi, \rho, \eta)$ are the parameters described above.

The probability that a specific allele is not observed is then

$$\mathrm{P}(Z_a^m = 0 \,|\, \mathbf{n}) = G\{C; \rho D_a^m(\phi, \xi, \mathbf{n}), \eta\}. \tag{2.9}$$

Considering a DNA sample of known composition, where it is known that the allele a is present in the sample, this also denotes the dropout probability.

As mentioned in Section 1.1.4, alleles can be non-detectable for reasons that are unrelated to the amount of template DNA; the allele can be out of range, or a mutation in the primer-binding site can result in the allele not being amplified at all. In such cases it could be difficult to explain the lack of a peak by the general variability of the gamma distribution (or it would need a higher threshold to be applied). We return to the issue of modelling explicitly such alleles in Section 4.5.

Dropout model for a sample with a single known contributor

Allelic dropout for single source traces was studied by Tvedebrink et al. (2009, 2012a), who fitted a logistic regression model to the dropout probability as

$$P(Z_a = 0 \,|\, \bar{h}) = \frac{\alpha \bar{h}^\beta}{1 + \alpha \bar{h}^\beta} \tag{2.10}$$

where $\bar{h}$ is an average of observed peak heights above threshold calculated across all markers. They found that one could use the same value of $\beta = -4.35$ for all

markers, whereas α was marker dependent. The independent variable $\bar{h}$ was used as a proxy for total amount of DNA.

In our model, the theoretical mean heterozygote peak height $\mu = \rho\eta$ would be a similar proxy for the total amount of DNA. To compare the two models, we assume that $\bar{h}$ approximately corresponds to the mean peak height μ. Thus our model has

$$P(Z_a = 0 \,|\, \mu) = G\{C; \mu/\eta, \eta\}, \tag{2.11}$$

which can then be compared with (2.10) where $\bar{h}$ is replaced by μ.

Figure 2.3 displays a selection of curves (2.11) for $C = 50$ and values of η corresponding to the maximum likelihood estimate and upper and lower 99% confidence limits obtained in the example analysis of Cowell et al. (2015); curves (2.10) for $\beta = -4.35$ and representative values of α are superimposed. We note that the dropout model based on the gamma distribution tends to have lower dropout rates than the logistic model for small amounts of DNA.

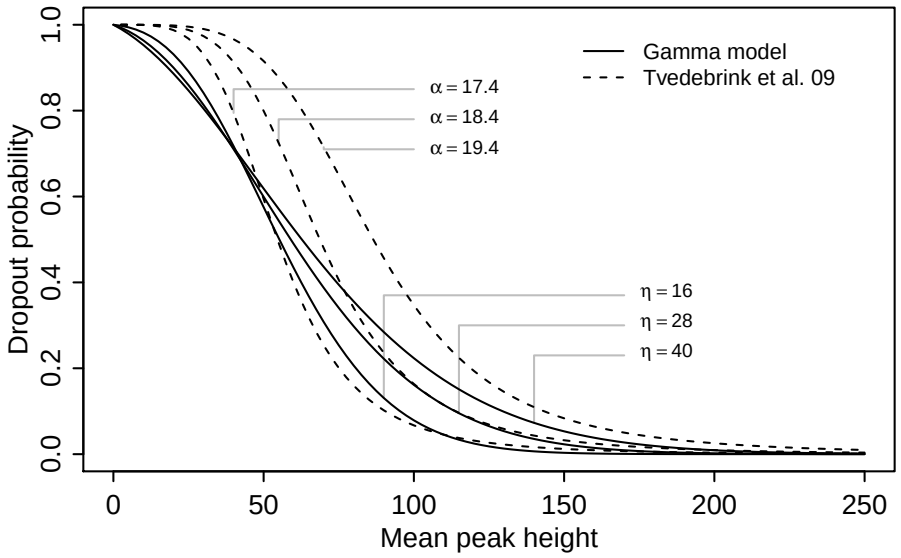


Figure 2.3: The probability of dropout of a single allele as a function of mean peak height. The curves with full lines correspond to our gamma model whereas the dashed curves correspond to the logistic model. The illustration is taken from Cowell et al. (2015).

The regression model of Tvedebrink et al. (2009) models dropout probability conditionally on the unknown peak heights. Balding (2013) adopts the logistic

model, although extends it to a relationship directly between the dropout probability and the amount of DNA.

2.2.4 Dropin

We take the view that all allelic peaks – including dropin – are explained by the presence of DNA from a person. Thus, dropin is modelled by including as many unknown contributors as needed to explain the mixture.

Due to the increase in complexity when including an additional unknown contributor, Balding (2013) models dropin as independent Bernoulli events. Puch-Solis et al. (2013) take the opposite view and prefer a dropin component, but in the lack of a such, they introduce additional unknown contributors. Steele and Balding (2014) highlight the difficulty in finding a suitable model for dropin and argue that using additional unknown contributors is therefore perhaps preferable.

In the event of a single spurious allele with a high peak, it would be unlikely that all other alleles of that person have dropped out and so we would expect the diagnostic methods, to be introduced in Chapter 6, to reveal that the peak is not explained well by the presence of an additional contributor.

2.2.5 Summary

The model for peak heights developed over the past few sections to allow for stutter and dropout may be summarised as follows.

Model 2.2: The peak height model in the potential presence of artefacts

Consider a mixture of DNA from individuals $i = 1, \ldots, k$ analysed for a set of markers $m = 1, \ldots, M$, each with a range of alleles $a = 1, \ldots, A_m$. Given the full set of genotypes, here represented by the set of allele counts

$$\boldsymbol{n} = \{n_{ia}^m; i = 1, \ldots, k; m = 1, \ldots, M; a = 1, \ldots, A_m\},$$

the set of peak heights

$$\{H_a^m; m = 1, \ldots, M; a = 1, \ldots, A_m\}$$

in the EPG are mutually independent and gamma distributed as

$$H_a^m \,|\, \boldsymbol{n} \sim \Gamma\left(\rho \sum_{i=1}^{k} \left\{(1-\xi) n_{ia}^m + \xi n_{i,a+1}^m\right\} \phi_i, \eta\right),$$

defining $n_{i,A_m+1}^m = 0$. Here

- k is the number of contributors
- n_{ia}^m is the number of alleles a that individual i possesses.
- ϕ_i is the proportion of DNA originating from individual i.
- ξ is the mean stutter proportion.
- ρ is proportional to the amount of DNA in the sample.
- η is the scale parameter.

A height threshold $C \geq 0$ is applied, and so only censored peak heights

$$\{Z_a^m; m = 1, \ldots, M; a = 1, \ldots, A_m\}$$

with

$$Z_a^m = \begin{cases} H_a^m, & H_a^m \geq C \\ 0, & H_a^m < C \end{cases}$$

are observable.

The basic gamma model (Model 2.1) is a special case of the peak height model with artefacts (Model 2.2), corresponding to $\xi = 0$ and also $C = 0$, so that no peak height threshold is used.

It should be noted that the independence of peak heights, and thus also that of censored peak heights, only holds conditionally on the genotypes of all contributors. As the genotypes are assumed independent between markers the peak heights are, however, independent between markers. A Bayesian approach integrating out the

parameters (ϕ, ρ, η, ξ), which are here considered fixed, will introduce dependence between the peak heights both within and between markers.

2.3 Likelihood function

The likelihood function is of direct interest in DNA mixture analysis, as it is used for assessing the evidential value of a stain. In addition, we shall use the likelihood function for parameter estimation by maximum likelihood. The set of peak heights $\{z_a^m,\ m = 1, \ldots, M;\ a = 1, \ldots, A_m\}$ constitutes the observed data and thus determines the likelihood function. Of further interest is the evaluation of the conditional distribution of the genotypes of unknown individuals given the observed peak heights. We let in the following $\psi = (\rho, \eta, \xi, \phi)$ denote the set of model parameters.

As the peak heights are conditionally independent given genotypes, the likelihood function is for fixed genotypes simply a product of the densities (2.8)

$$\begin{aligned} f_\psi(z_a^m;\ m = 1, \ldots, M;\ a = 1, \ldots, A_m \,|\, \boldsymbol{n}, H) &= \prod_{m=1}^{M} \prod_{a=1}^{A_m} f_\psi(z_a^m \,|\, \boldsymbol{n}, H) \\ &= \prod_{m=1}^{M} \prod_{a=1}^{A_m} f_\psi(z_a^m \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}, H). \end{aligned}$$

The full likelihood function is the marginal density of the peak heights

$$\begin{aligned} L(\psi \,|\, H) &= f_\psi(z_a^m; m = 1, \ldots, M;\ a = 1, \ldots, A_m \,|\, H) \\ &= \prod_{m=1}^{M} f_\psi(z_1^m, \ldots, z_{A_m}^m \,|\, H) \\ &= \prod_{m=1}^{M} \mathbb{E}\Big\{ f_\psi\Big(z_1^m, \ldots, z_{A_m}^m \,\Big|\, \boldsymbol{n}^m\Big) \,\Big|\, H \Big\} \\ &= \prod_{m=1}^{M} \mathbb{E}\Big\{ \prod_{a=1}^{A_m} f_\psi(z_a^m \,|\, \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m) \,\Big|\, H \Big\}. \end{aligned} \tag{2.12}$$

Here the expectations are taken with respect to the distribution of genotypes $\boldsymbol{n}^m$, $m = 1, \ldots, M$ of the contributors under the hypothesis H under consideration.

The expectation in (2.12) is the sum over all configurations of possible genotypes of the contributors under H as

$$\mathbb{E}\left\{\prod_{a=1}^{A_m} f_\psi(z_a^m \mid \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m) \,\middle|\, H\right\} = \sum_{\boldsymbol{n}^m} \left\{\prod_{a=1}^{A_m} f_\psi(z_a^m \mid \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m)\right\} p(\boldsymbol{n}^m \mid H).$$

Direct computation of this sum is typically infeasible when there are many alleles and many unknown contributors. The sum has $\{A_m(A_m+1)/2\}^k$ terms, as this is the number of possible combinations of the k genotypes at a marker with A_m possible alleles; furthermore, each term is a product of A_m factors corresponding to the densities for the peak heights at the A_m allelic positions in the EPG for marker m. In Chapter 4 we discuss the computational complexity of various approaches for computing the expectation (2.12).

2.4 A joint model for multiple mixtures

Sometimes it is of interest to make a joint model for the peak heights across a set of observed EPGs. For instance, we may have *replicate analyses* of a single DNA sample, so that one EPG is produced for each of multiple extracts of one DNA sample; in this case the composition of the samples are roughly the same. In other instances, we may have available multiple mixtures with potentially different sets of donors, but perhaps the mixtures share a connection to a particular crime case, or they have been analysed under the same laboratory settings.

Whatever the specific scenario, one reason for modelling the DNA samples jointly is to exploit the gain in information about any contributors and model parameters shared between the samples. Another reason is to produce a single evidential value for the entire set of DNA mixtures; simply multiplying the likelihood ratios computed separately for each mixture does not take into account the dependence when mixtures share contributors.

It is natural to assume that EPGs are observed independently, given the genotypes of any shared contributors. Multiplying the likelihood ratios then corresponds to a model in which each DNA sample has its separate set of contributors, independent of those of the other samples.

Consider the joint set of contributors $i = 1, \ldots, k$, where contributor i have contributed to EPG e with a proportion ϕ_{ei}, allowing $\phi_{ei} = 0$ and thus the possibility that contributor i is not an actual donor to the DNA sample e.

Where $R \geq 1$ mixtures are modelled jointly, we have one set of model parameters for each mixture, and so the total set of parameters under the joint model is

$$\begin{array}{c} \\ 1 \\ \vdots \\ R \end{array} \begin{array}{c} \begin{array}{cccc} \rho & \eta & \xi & \phi \end{array} \\ \begin{pmatrix} \rho_1 & \eta_1 & \xi_1 & \phi_{11}, \ldots, \phi_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_R & \eta_R & \xi_R & \phi_{R1}, \ldots, \phi_{Rk} \end{pmatrix} \end{array}.$$

Under the joint model, the unknown contributors are ordered according to their contributions to the first mixture only, because the order of the contributions from shared contributors may not be the same across mixtures. In the event that the first mixture is unable to distinguish the contributions from two unknown contributors, additional order constraints may need to be imposed to ensure identifiability, but we ignore this here.

The joint model is summarised in Model 2.3 below. Note that we allow the set of markers to differ between EPGs; this is relevant, for instance, when different STR marker systems are used for the analysis of the samples.

Model 2.3: Peak height model for multiple mixtures

Let H_{ea}^m denote the peak height for marker m, allele a in the EPG e. We model the total set of peak heights

$$\{H_{ea}^m; e = 1, \ldots, R; m = 1, \ldots, M_e; a = 1, \ldots, A_m\}.$$

Conditionally on the genotypes $\boldsymbol{n}$ for the joint set of contributors to the EPGs, the peak heights are mutually independent and distributed as

$$H_{ea}^m \sim \Gamma\left(\rho_e \sum_{i=1}^{k} \left\{\xi_e n_{ia}^m + (1-\xi_e) n_{i,a+1}^m\right\} \phi_{ei}, \eta_e\right)$$

Applying a detection threshold $C_e \geq 0$ for each EPG, the observed peak heights are

$$Z_{ea}^m = \begin{cases} H_{ea}^m, & H_{ea}^m \geq C_e \\ 0, & H_{ea}^m < C_e. \end{cases}$$

Note that Model 2.2 for a single mixture arises as the special case $R = 1$.

2.5 A qualitative model

Rather than exploiting the full peak height information for the interpretation of a DNA mixture, it may at times be of interest to use a more simplistic model. Our quantitative model for DNA mixtures induces a qualitative model, in that the peak height distribution implies a distribution for the presence ($Z_a^m \geq C$) or absence ($Z_a^m = 0$) of peaks in the EPG. Under this model, rather than observing peak heights $\{z_a^m,\ m = 1, \ldots, M;\ a = 1, \ldots, A_m\}$ across markers and alleles, we observe only whether or not a peak has been seen,

$$\{\mathbb{1}_{\{z_a^m \geq C\}};\ m = 1, \ldots, M;\ a = 1, \ldots, A_m\}.$$

As the peak heights Z_a^m are mutually independent given genotypes, so are the dichotomous peak-presence variables $\mathbb{1}_{\{z_a^m \geq C\}}$. Further, conditionally on the genotypes, their distribution is given by

$$p_\psi(\mathbb{1}_{\{z_a^m \geq C\}} = i \,|\, \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m) = \begin{cases} G_\psi\Big(C \,\Big|\, \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m\Big), & i = 0 \\ 1 - G_\psi\Big(C \,\Big|\, \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m\Big), & i = 1. \end{cases} \tag{2.13}$$

Analogously to the likelihood function based on peak heights, the likelihood function for the discrete model is

$$\tilde{L}(\psi \mid H) = \prod_{m=1}^{M} \mathbb{E}\left(\prod_{a=1}^{A_m} p_\psi(\mathbb{1}_{\{z_a^m \geq C\}} \mid \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m) \,\middle|\, H \right). \tag{2.14}$$

Although less informative, the more simplistic model avoids some of the assumptions needed in modelling the full distribution of peak heights and may therefore be more robust. Another use is for the comparison to discrete models that are currently applied in casework.

Chapter 3

Bayesian network techniques

The need for marginalisation over the genotypes of unknown contributors in computations involving the peak-height distribution typically requires a significant computational effort. Our computational approach to performing the required marginalisation relies on Bayesian network techniques.

A brief introduction to standard techniques for Bayesian networks is given in Section 3.1 below, focusing on the Bayesian network as a computational device. A more detailed introduction may be found in e.g. Cowell et al. (1999). The exposition of probability propagation in Section 3.1 largely follows the exposition of Dawid (1992) as described in Cowell et al. (1999, Section 6.3).

In Section 3.2 we then introduce a new technique for computing the expectation of a product of random variables through probability propagation in a Bayesian network augmented with appropriate auxiliary variables (Graversen and Lauritzen, 2014). The relevance of computing such expectations is directly motivated by the need for marginalisation, i.e. take expectation, over the space of DNA profiles in computing the likelihood function (2.12); we return to this specific application in Chapter 4.

3.1 An introduction to inference in Bayesian networks

Let $X = (X_v)_{v \in V}$ be a collection of discrete random variables each with a finite state space $\mathcal{X}_v$. We are interested in representations of the distribution of X, specifically factorisations of its probability mass function $p(x)$ over subsets of the variables. For further inference in the model, we seek a representation that allows easy computation of various marginal and conditional distributions of the variables.

3.1.1 Bayesian network representation

A *Bayesian network* represents the distribution of $X = (X_v)_{v \in V}$ by a *directed acyclic graph (DAG)* with node set V; an example is shown in Figure 3.1. Each node $v \in V$ corresponds to the random variable X_v, and the terms "random variable" in the model and "node" in the DAG are in practice used interchangeably. The set of edges in the DAG specify a factorisation of $p(x)$ according to subsets of the variables,

$$p(x) = \prod_{v \in V} p(x_v | x_{\mathrm{pa}(v)}). \tag{3.1}$$

Here, $\mathrm{pa}(v)$ denotes the set of parents of the node v and, for $B \subseteq V$, X_B denotes the subset $(X_v)_{v \in B}$ of the variables. For a node with no parents, $p(x_v | x_{\mathrm{pa}(v)}) = p(x_v)$.

Thus, a Bayesian network for the distribution of X is the combination of a DAG and the collection of conditional probability mass functions $p(x_v | x_{\mathrm{pa}(v)})$ of any variable given its parents, as specified by the DAG. Of importance for further computations, each of these factors depends only on a subset of the variables, and $p(x)$ may be found as their product.

Example 3.1. Any distribution can be represented by a Bayesian network, noting that

$$p(x_1, \ldots, x_n) = \prod_{v=1}^{n} p(x_v \,|\, x_1, \ldots, x_{v-1})$$

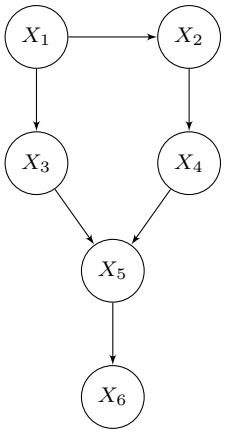


Figure 3.1: A DAG. The variable X_5 has *parents* X_3 and X_4; X_5 is their *child*. Further, X_6 is a *descendant* of X_5 and indeed also of all other variables in the model. The variables X_1, X_2, and the parents of X_5 are then the *non-descendants* of X_5.

factorises according to a DAG where any node X_v has parents $X_1, \ldots, X_{v-1}$. However, it is difficult to give a general description of how to find a good Bayesian network for a given set of variables. As we shall see below, a factorisation over small sets of variables is generally preferable, and in the above case the n'th factor depends on all of the variables. •

The density $p(x)$ factorises according to the DAG as (3.1) exactly when the *local directed Markov property* holds, i.e. when any variable is conditionally independent of its non-descendants given the set of parent variables. This provides a way of proving that a particular Bayesian network is indeed a correct specification of the distribution of interest. In Figure 3.1, the local directed Markov property gives, for instance, that X_5 is conditionally independent of X_1 and X_2 given X_3 and X_4; this is written $X_5 \perp\!\!\!\perp X_1, X_2 \,|\, X_3, X_4$.

3.1.2 Junction tree representation

When a model is specified in terms of a Bayesian network, the further inference is performed on a more convenient computational structure, a so-called *junction tree*. As for the Bayesian network, the distribution is represented by the combination of

a graph – the junction tree – and a collection of functions of subsets of variables, from which $p(x)$ can be computed.

The tree structure. In a junction tree for the distribution of X, the node set $\mathcal{C}$ consists of subsets of V, with each node $C \in \mathcal{C}$ corresponding to a subset X_C of variables. The nodes of a junction tree are called *cliques*, as they correspond to maximal complete subsets – cliques – in an associated graph, which we introduce below. The junction tree is itself a tree with the additional and characterising property that for any $v \in V$ the set of cliques C such that $v \in C$ induces a subtree, i.e. a connected subgraph. Figure 3.2d shows a junction tree representing the six variables modelled by the DAG in Figure 3.1.

For any two neighbouring cliques C_1 and C_2, we associate to the edge between them a *separator* $C_1 \cap C_2$, which then corresponds to the set $X_{C_1 \cap C_2}$ of variables that the two cliques have in common. We denote by $\mathcal{S}$ the set of separators indexed by the edges of the junction tree, so that if a separator is associated to multiple edges it occurs multiple times in $\mathcal{S}$.

The factorisation of $p(x)$. The distribution of X is represented by an unnormalised probability mass function $g(x)$ of the form

$$p(x) \propto g(x) = \frac{\prod_{C \in \mathcal{C}} \zeta_C(x_C)}{\prod_{S \in \mathcal{S}} \zeta_S(x_S)}, \tag{3.2}$$

where ζ_C and ζ_S are non-negative functions called *potentials*, and the collection of potentials is the *charge*. The normalising constant needed to obtain $p(x)$ is $\sum_x g(x)$.

Cutting the junction tree along an edge with associated separator S results in two disconnected subtrees partitioning V into two sets of nodes, V_1 and V_2, with $V_1 \cap V_2 = S$. This is a consequence of the junction tree property; for any $v \in V_1 \cap V_2$, the subtree spanned by the cliques containing v will necessarily include both the removed edge and the two cliques that were originally connected by it. Thus v lies in both of these cliques and therefore also in their intersection, the separator S.

An important property of the junction tree representation – a consequence of the factorisation (3.2) – is that the variables X_{V_1} and X_{V_2} are conditionally independent given their shared variables X_S.

The marginal charge. The potentials in (3.2) can be chosen in many ways for a given function g. Given a junction tree and an arbitrary charge, a message-passing operation referred to as *propagation* brings the charge to a canonical form; the propagation algorithm is discussed below. Each of the potentials in the canonical charge – the *marginal charge* – are equal to the function g marginalised onto the corresponding clique or separator, i.e. for all cliques or separators D

$$\zeta_D(x_D) = \sum_{y:y_D=x_D} g(y).$$

Thus, in the marginal charge the potential ζ_D is the (unnormalised) marginal probability mass function for X_D.

The normalising constant for $g(x)$ is clearly the same as the normalising constant for any of the potentials in the marginal charge and is therefore efficiently computed as

$$\sum_x g(x) = \sum_{x_D} \zeta_D(x_D), \tag{3.3}$$

for any clique or separator D. A good choice of D for computing the normalising constant is a separator $D \in \mathcal{S}$ with minimal state space.

Marginal distributions

An advantage of using the junction tree in combination with the marginal charge is that it facilitates the further computation of various marginal distributions. Evidently, the distributions of $X_C, C \in \mathcal{C}$ and $X_S, S \in \mathcal{S}$ are readily available from the potentials.

More generally, consider a subtree of the junction tree, and let the charge on this – smaller – junction tree consisting of a set of cliques $\mathcal{C}' \subseteq \mathcal{C}$ and separators

$\mathcal{S}' \subseteq \mathcal{S}$ be formed of the corresponding collection of potentials from the junction tree representing $g(x)$. The new junction tree covers the nodes $V' = \bigcup_{C \in \mathcal{C}'} C$, and it holds that

$$p(x_{V'}) \propto g(x_{V'}) = \frac{\prod_{C \in \mathcal{C}'} \zeta_C(x_C)}{\prod_{S \in \mathcal{S}'} \zeta_S(x_S)}, \tag{3.4}$$

meaning that the junction tree is a correct representation of the distribution of $X_{V'}$. The marginal charge is the only charge satisfying (3.4), making it a particularly suitable choice as a canonical form.

As a consequence, the marginal distribution of any set of variables can be obtained by "local" marginalisation over variables in the smaller junction tree rather than over the entire state space: For any set of variables X_A, we find a subtree with cliques $\mathcal{C}'$ such that $A \subseteq V' = \bigcup_{C \in \mathcal{C}'} C$ and then compute

$$p(x_A) = \sum_{\substack{y \in \mathcal{X}_V \\ y_A = x_A}} p(y) = \sum_{\substack{y \in \mathcal{X}_{V'} \\ y_A = x_A}} p(y_{V'}),$$

were $\mathcal{X}_B$ denotes the state space $\bigtimes_{v \in B} \mathcal{X}_v$ of X_B. The distribution of a single variable is most easily computed directly from the potential on a minimal separator containing the variable of interest.

Conditional distributions

Another advantage of the junction tree representation is that the same tree can be used for the representation of various conditional distributions.

The function $g(x)$ represented by a junction tree can be modified by *entering likelihood evidence* $\ell_v(x_v)$, $v \in V$, denoting the process of multiplying $g(x)$ by non-negative functions $\ell_v(x_v)$. The multiplication takes place at the level of individual potentials, so that the factor $\ell_v(x_v)$ is multiplied onto the potential $\zeta_C(x_C)$ for some clique C containing v. The junction tree together with the modified charge then represents

$$\tilde{g}(x) = g(x) \prod_{v \in V} \ell_v(x_v)$$

with normalising constant

$$\sum_x \tilde{g}(x) = \sum_x g(x) \prod_{v \in V} \ell_v(x_v). \tag{3.5}$$

By a subsequent propagation, the modified charge is brought to its canonical form, where all potentials are the marginals of $\tilde{g}$.

As an immediate consequence, a representation of the conditional distribution of X given a subset $X_A = \xi_A$ of the variables can be obtained by for $v \in A$ propagating likelihood evidence $\ell_v(x_v) = \mathbb{1}_{\{x_v = \xi_v\}}$ in the junction tree representing the distribution of X.

Further, for a – possibly continuous – random variable Y that is not in the network and only depends on X through a single node X_v, a junction tree representation of the conditional distribution of X given $Y = y$ is obtained by propagating likelihood evidence $\ell_v(x_v) = f(y \,|\, x_v)$ based on the density for the conditional distribution of Y given X.

This way of incorporating a continuous variable in an otherwise discrete model represented by a Bayesian network has been exploited in the setting of DNA mixtures, e.g. in Cowell et al. (2007a) and Graversen and Lauritzen (2013), to obtain a representation of the conditional distribution of contributors' (discrete) genotypes given the (continuous) observed peak heights.

Setting up a junction tree

We now describe how to obtain a junction tree representation of a model, when a Bayesian network is already available.

Triangulation. Any distribution represented by a Bayesian network can also be represented by a factorisation over the cliques – maximal complete subsets – in a *chordal* graph, i.e. an undirected graph where in any cycle of four or more distinct

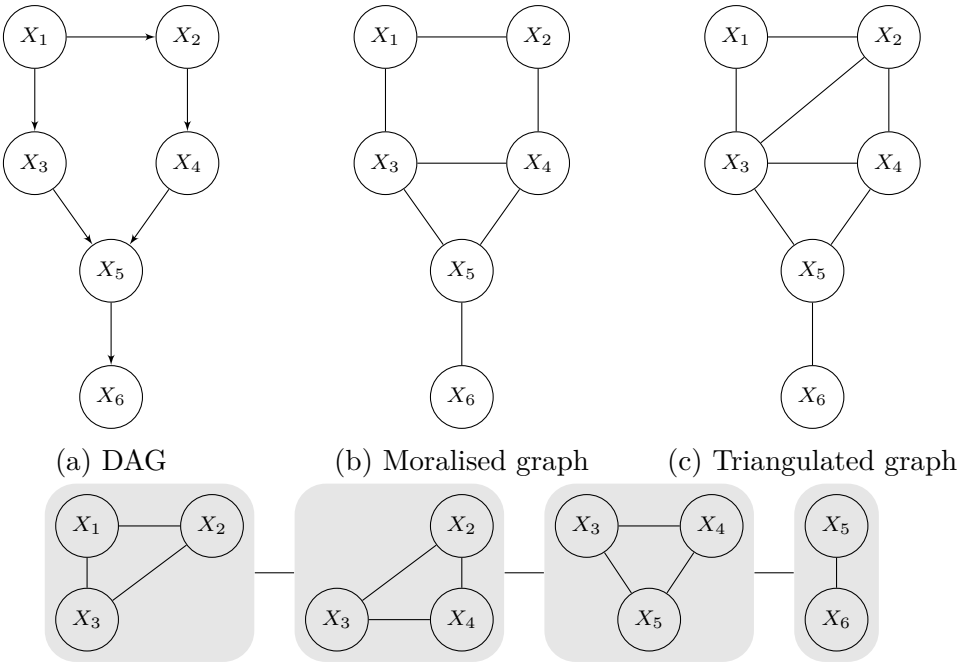
nodes there are two non-consecutive nodes with an edge (a chord) between them. Important properties for the representation by a chordal graph include that its cliques can always be arranged into a junction tree, and that the potentials in the factorisation (3.2) over cliques and separators can be chosen to be the marginal charge.

The process of obtaining a chordal graph from a Bayesian network is done in two steps.

The first step – the *moralisation* – is to create an undirected graph, in which the sets $\{v\} \cup \mathrm{pa}(v)$ are all complete and hence each contained in at least one clique. By definition of the Bayesian network, $p(x)$ factorises according to the node sets $\{v\} \cup \mathrm{pa}(v)$ and thus also according to cliques in the moralised graph. The construction of the moralised graph is straight-forward: An undirected edge is added between any two nodes that have a common child and any existing directed edge is converted to an undirected edge (Figure 3.2b).

If the moralised graph is not itself chordal, the second step – the *triangulation* – is to add enough edges that the resulting graph is chordal (Figure 3.2c). As adding edges can only result in larger cliques, the factorisation of $p(x)$ over cliques is preserved. The triangulation can generally be done in many ways; finding an optimal triangulation for a given optimality criterion is an NP-hard problem (Yannakakis, 1981). In Figure 3.2c, adding the edge X_1—X_4 rather than the edge X_2—X_3 or even adding both these edges would also have resulted in a chordal graph, but the corresponding set of cliques would have been different.

From triangulated graph to junction tree. Identification of the cliques in the triangulated graph and their further arrangement into a junction tree can be done by standard machinery, for instance based on maximum cardinality search. The cliques can often be arranged in multiple ways to form a junction tree, but note that although a separator is determined by the intersection of two neighbouring



(a) DAG (b) Moralised graph (c) Triangulated graph

(d) Junction tree of cliques

Figure 3.2: Various graphical representations of a distribution

cliques, the set of separators is the same regardless of the choice of junction tree. As the specific choice of junction tree is unimportant for our purposes, we leave out a description of this process.

Initial charge. One way of obtaining a representation (3.2) of $p(x)$ is as follows. For any separator $S \in \mathcal{S}$, let $\zeta_S(x_S) = 1$. For the clique potentials, assign each node v to some clique $C \in \mathcal{C}$, in which v is contained. If no nodes are assigned to clique C, we set $\zeta_C(x_C) = 1$. Otherwise, denoting by $V(C)$ the set of nodes assigned to a clique C, we set

$$\zeta_C(x_C) = \prod_{v \in V(C)} p(x_v | x_{\mathrm{pa}(v)}).$$

3.1.3 Local computations on a junction tree

Probability propagation

As mentioned, the potentials in a charge are not unique. New representations can be generated from an existing charge by a message-passing operation modifying pairs

of potentials. Importantly, the marginal charge can be reached by a sequence of such operations and the resulting algorithm is called propagation.

To *pass a message* along an edge S' from a clique C to a neighbouring clique C', denotes the modification of the potentials $\zeta_{S'}$ and $\zeta_{C'}$ to

$$\zeta^*_{S'}(x_{S'}) = \sum_{\substack{y \in \mathcal{X}_C \\ y_{S'} = x_{S'}}} \zeta_C(y_C) \qquad \text{and} \qquad \zeta^*_{C'}(x_{C'}) = \frac{\zeta^*_{S'}(x_{S'})}{\zeta_{S'}(x_{S'})} \zeta_{C'}(x_{C'}).$$

Since $\zeta^*_{C'}/\zeta^*_{S'} = \zeta_{C'}/\zeta_{S'}$, is clear that the product (3.2) for computing $g(x)$ from the potentials is invariant to the message-passing operation.

In the propagation algorithm used in HUGIN, messages are passed along each edge twice. Choosing any clique as the root, imagine for the purpose of illustration that we direct all edges in the junction tree away from the root. The result is a DAG (of cliques), in which any clique has at most one parent and possibly multiple children. In the first of two phases, messages are passed along all edges against the directions and toward the root clique; once a clique has received messages from its children, if any, it passes a message to its parent. In the second phase, messages are again passed along all edges outward from the root in the direction of the edges; once a clique has received a message from its parent, it passes a message to its children. After the two phases, the charge is in canonical form, the marginal charge.

Total size of the junction tree

In applications where there is a high number of states for X, time- and memory-consumption for the junction tree and associated algorithms is an important issue.

If functions are stored as look-up tables, then storing the probability mass function p directly requires a cell per configuration of X; there are $\prod_{v \in V} |\mathcal{X}_v|$ such configurations.

Alternatively, we can store only a representation of p in terms of the separator- and clique-potentials in a charge. The table for a single potential on a clique or

separator D contains $\prod_{v \in D} |\mathcal{X}_v|$ real numbers and so we need a table of

$$\sum_{C \in \mathcal{C}} \prod_{v \in C} |\mathcal{X}_v| + \sum_{S \in \mathcal{S}} \prod_{v \in S} |\mathcal{X}_v|$$

numbers for the full representation of p. This quantity is called the *total size* of the junction tree. It is clear that in terms of memory consumption, a junction tree of small total size is preferable.

For the probability propagation, where messages are passed along each edge twice, all tables associated with cliques and separators need to be updated twice in the worst case. Thus, the number of elementary arithmetic operations for propagation is linear in the total size. Again, a small total size is preferable.

Simulation

Simulation from a Bayesian network is simple; variables can be sampled sequentially using any order that respects that a particular variable is sampled only after a configuration of its parents has already been sampled, since then it can be used that the variable is independent of its non-descendants given this configuration of the parents.

The junction tree representation is a more flexible representation than the DAG in that it is easily obtained for various conditional distributions; indeed the structure of the graph is maintained and only a modification of the potentials is required.

Further, the junction tree allows a sampling scheme similar to that of the DAG, perhaps well illustrated by thinking again of the junction tree as inducing a DAG of cliques, from which we can then sample. In this analogy, each clique has a single parent with the exception of the root clique, which has none. The distribution of X_C given its non-descendants – among which are the already visited cliques – only depends on the separator between X_C and its parent clique and is given by $p_C(X_C)/p_S(X_S)$, which is readily available through the marginal charge.

3.2 Expectations of products using auxiliary variables

We have seen in Section 3.1 that the normalising constant for the unnormalised probability function on the junction tree may easily be computed after propagation, using (3.3). This provides a way of computing a sum over the entire state space of the Bayesian network. In this section, we shall build on this idea and develop a general technique for computing expectations. The exposition of our new methodology follows closely that of Graversen and Lauritzen (2014).

Consider a collection $X = \{X_v\}_{v \in V}$ of discrete variables with a distribution represented by a Bayesian network, and let h be a non-negative function that can be factorised as

$$h(x) = \prod_{B \in \mathcal{B}} h_B(x_B), \tag{3.6}$$

for some set $\mathcal{B}$ of subsets of V and real-valued non-negative functions h_B. In the following, we propose a general technique for computing $\mathbb{E}\, h(X)$.

3.2.1 Summation by propagation

Let the junction tree represent the unnormalised probability mass function $g(x)$ for the distribution of X, and let $N_1 = \sum_x g(x)$ be the corresponding normalising constant. Similarly, let $N_2 = \sum_x g(x) \prod_{v \in V} \ell_v(x_v)$ be the normalising constant after propagating likelihood evidence $\prod_{v \in V} \ell_v(x_v)$. Taking the ratio of the normalising constants N_1 and N_2 yields the expectation of the product of the likelihood evidence

with respect to $p(x) = g(x)/\sum_x g(x)$:

$$\begin{aligned}\frac{N_2}{N_1} &= \frac{\sum_x g(x) \prod_{v\in V} \ell_v(x_v)}{\sum_y g(y)} \\ &= \sum_x \frac{g(x)}{\sum_y g(y)} \prod_{v\in V} \ell_v(x_v) \\ &= \sum_x p(x) \prod_{v\in V} \ell_v(x_v) \\ &= \mathbb{E}\left\{ \prod_{v\in V} \ell_v(X_v) \right\}. \end{aligned} \tag{3.7}$$

In situations where the function h factorises over nodes in the network we may thus simply compute its expectation $\mathbb{E}\, h(X)$ by entering each factor in the product as likelihood evidence for the corresponding node; this method was used in Graversen and Lauritzen (2013) for computing the likelihood function for relative peak heights in the model of Cowell et al. (2007a). However, we are interested in a more general framework, where h may factorise over arbitrary subsets of nodes.

An important use of (3.7) is for the computation of the joint probability for a specific configuration $X_{V'} = \xi_{V'}$ of a subset of network variables. Defining $\ell_v(x_v) = \mathbb{1}_{\{x_v=\xi_v\}}$, the ratio (3.7) of normalising constants yields

$$\mathbb{E}\left(\prod_{v\in V} \ell_v(X_v) \right) = \mathbb{E}\left(\prod_{v\in V'} \mathbb{1}_{\{X_v=x_v\}} \right) = \mathrm{P}\left(\bigcap_{v\in V'} \{X_v = x_v\} \right). \tag{3.8}$$

3.2.2 Constructing auxiliary variables

For the purpose of computing $\mathbb{E}\, h(x) = \mathbb{E} \prod_{B\in\mathcal{B}} h_B(x_B)$, we introduce for each $B \in \mathcal{B}$ a binary random variable $Y^B \in \{0, 1\}$. These *auxiliary* variables are assumed to be mutually conditionally independent given the network and distributed as

$$\mathrm{P}\Big(Y^B = 1 \,\Big|\, X = x\Big) = \mathrm{P}\Big(Y^B = 1 \,\Big|\, X_B = x_B\Big) = h_B(x_B)/k_B. \tag{3.9}$$

Here, the constant k_B is chosen such that $h_B(x_B)/k_B \in [0, 1]$ over all states x_B and so (3.9) defines a valid probability distribution. A simple choice would be $k_B = \max_{x_B} h_B(x_B)$, i.e. the largest value that h_B attains over the state space of

X_B. We use the state space $\{0, 1\}$ for auxiliary variables, but note that this choice is unimportant for the method itself.

The desired expectation $\mathbb{E}\{\prod_{B\in\mathcal{B}} h_B(X_B)\}$ can now be expressed in terms of the probability of a specific configuration of the auxiliary variables introduced. As Lemma 1 reveals, this is also the case for the expectation of a product of any subset of the variables $h_B(X_B)$.

Lemma 1. *For all $\mathcal{B}' \subseteq \mathcal{B}$ it holds that*

$$\mathbb{E}\left\{\prod_{B\in\mathcal{B}'} h_B(X_B)\right\} = \mathrm{P}\left(\bigcap_{B\in\mathcal{B}'} \{Y^B = 1\}\right) \prod_{B\in\mathcal{B}'} k_B.$$

Proof. Using (3.9) and the fact that $Y^B, B \in \mathcal{B}$ are mutually conditionally independent given X we get

$$\begin{aligned}
\mathbb{E}\left\{\prod_{B\in\mathcal{B}'} h_B(X_B)\right\} &= \mathbb{E}\left\{\prod_{B\in\mathcal{B}'} \Big(\mathrm{P}\big(Y^B = 1 \,\big|\, X_B\big) k_B\Big)\right\} \\
&= \mathbb{E}\left\{\prod_{B\in\mathcal{B}'} \mathrm{P}\big(Y^B = 1 \,\big|\, X\big)\right\} \prod_{B\in\mathcal{B}'} k_B \\
&= \mathbb{E}\left\{\mathrm{P}\Big(\bigcap_{B\in\mathcal{B}'} \{Y^B = 1\} \,\Big|\, X\Big)\right\} \prod_{B\in\mathcal{B}'} k_B \\
&= \mathrm{P}\left(\bigcap_{B\in\mathcal{B}'} \{Y^B = 1\}\right) \prod_{B\in\mathcal{B}'} k_B
\end{aligned}$$

as desired. □

Lemma 1 allows the interpretation of the expectation of interest as a scaled probability, which can be computed in various ways; Proposition 1 below provides one such way.

Extending the Bayesian network

The Bayesian network representing the distribution of $\{X_v\}_{v\in V}$ can be extended to include the variables $\{Y^B\}_{B\in\mathcal{B}}$ by for each B adding Y^B as a child of $\{X_v\}_{v\in B}$ with conditional distributions of Y^B as given in (3.9). Because the auxiliary variables are added as children of existing network nodes, no directed cycles are created and the

extended network is a correct representation of the joint distribution of (X, Y) since, given X_B, Y^B is conditionally independent of all other variables in the extended network.

Figure 3.3 illustrates how the network is extended in case of a function h factorising over two sets of variables (X_2, X_3) and (X_3, X_4, X_5).

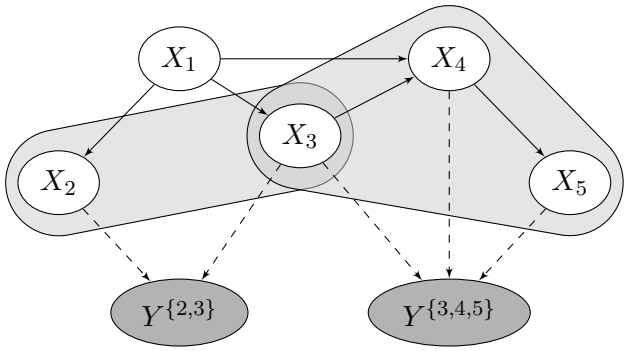


Figure 3.3: Extending a network with two binary variables for computation of $\mathbb{E}\left(h_{\{2,3\}}(X_2, X_3)h_{\{3,4,5\}}(X_3, X_4, X_5)\right)$. Here $\mathcal{B} = \{\{2, 3\}, \{3, 4, 5\}\}$

Note that as X_B is the parent set of Y^B in the extended network, the node set X_B will always be a complete set in the triangulated graph and thus contained in some clique.

As shown in Proposition 1 below, the property (3.7) ensures that the expectation of interest can be calculated by propagating likelihood evidence on the auxiliary variables. The computation in Proposition 1 is essentially equivalent to computing the probability in Lemma 1 by propagation as in (3.8), just that we in addition propagate the scaling factors and thus compute the scaled probability directly.

Proposition 1. *Let likelihood evidence for each node Y^B, $B \in \mathcal{B}' \subseteq \mathcal{B}$ be given as:*

$$\ell_B(Y^B) = \begin{cases} k_B, & Y^B = 1 \\ 0, & Y^B = 0 \end{cases}$$

and let N_1 and N_2 be the normalising constants before and after propagation of likelihood evidence. Then we have

$$\mathbb{E}\left\{\prod_{B\in\mathcal{B}'} h_B(X_B)\right\} = \frac{N_2}{N_1}.$$

Proof.

$$\begin{aligned}\frac{N_2}{N_1} &= \mathbb{E}\left\{\prod_{B\in\mathcal{B}} \ell_B(Y^B)\right\} \\ &= \mathbb{E}\left(\prod_{B\in\mathcal{B}'} k_B \mathbb{1}_{\{Y^B=1\}}\right) \\ &= \mathrm{P}\left(\bigcap_{B\in\mathcal{B}'} \{Y^B = 1\}\right) \prod_{B\in\mathcal{B}'} k_B\end{aligned}$$

which by Lemma 1 equals the desired expectation. □

Proposition 1 leads to a practical way of computing the desired expectation: for each auxiliary variable Y^B we compute the conditional probabilities $\mathrm{P}(Y^B \mid X_B)$ for all configurations of (Y^B, X_B), and thereafter the expectation can be obtained in a single propagation of the entered likelihood evidence.

Complexity considerations

In the worst case, the total size of the junction tree as defined in Section 3.1.3 determines the number of table cells that need updating when changing the conditional probabilities for auxiliary variables. The following propagation of likelihood evidence entered for the auxiliary variables is linear in the total size (Section 3.1.3). Thus, the efficiency of computation by auxiliary variables depends crucially on the total size of the junction tree.

Chapter 4

A Bayesian network representation of the DNA mixture model

In this chapter, we discuss how the joint model for peak heights and genotypes can be represented in terms of a Bayesian network. We propose to represent a single genotype by a Bayesian network that exhibits a Markovian structure and describe how this representation in combination with appropriate auxiliary variables yields a powerful and flexible Bayesian network representation of the joint model for peak heights and genotypes. In particular, the representation allows efficient evaluation of a wide range of quantities of interest in a case analysis. This chapter is largely based on Graversen and Lauritzen (2014).

In the following, we restrict attention to representations of the joint model $\Pr(\boldsymbol{Z}^m, \boldsymbol{n}^m \,|\, \psi, H)$ for peak heights and genotypes at a single locus. It should be noted that, under the standard model for DNA profiles (Section 2.1), there is independence between markers in the joint model $\Pr(\boldsymbol{Z}, \boldsymbol{n} \,|\, H)$ for DNA profiles and peak heights, and thus the corresponding networks are not connected. The independence

between markers follows from

$$\begin{aligned}\Pr(\boldsymbol{Z}, \boldsymbol{n} \,|\, \psi, H) &= \Pr(\boldsymbol{Z} \,|\, \boldsymbol{n}, \psi, H) \Pr(\boldsymbol{n} \,|\, H) \\ &= \prod_{m=1}^{M} \Pr(\boldsymbol{Z}^m \,|\, \boldsymbol{n}, \psi, H) \prod_{m=1}^{M} \Pr(\boldsymbol{n}^m \,|\, H) \qquad (4.1) \\ &= \prod_{m=1}^{M} \Pr(\boldsymbol{Z}^m \,|\, \boldsymbol{n}^m, \psi, H) \prod_{m=1}^{M} \Pr(\boldsymbol{n}^m \,|\, H). \qquad (4.2)\end{aligned}$$

Here (4.1) is due to the conditional independence of peak heights given genotypes and to the marginal independence of genotypes between markers. In (4.2) it is used that the peak heights at marker m only depend on the full set of genotypes through the genotypes for marker m.

4.1 Markov representation of a genotype

We here focus on modelling a single genotype at a specific marker; under the standard model for DNA profiles of unknown genotypes, Section 2.1, the genotypes are mutually independent both within and across individuals and so their networks are not connected.

The multinomial distribution of allele counts $(n_{i1}, \ldots, n_{iA})$ representing the genotype of an unknown contributor i does not in itself have Markovian properties. However, if we define the partial sums

$$S_{ia} = \sum_{b=1}^{a} n_{ib}$$

counting the number of alleles of type up to and including a that person i possesses, we can represent the genotype in a Bayesian network of the structure displayed in Figure 4.1.

If we imagine the two alleles in the genotype being allocated sequentially according to allele frequencies $(q_1, \ldots, q_A)$, then given the allocations of alleles of types $1, \ldots, a$ the distribution of the number of alleles that a person has of type $a+1$ only depends on how many alleles of the total two are left to allocate; further, the allo-

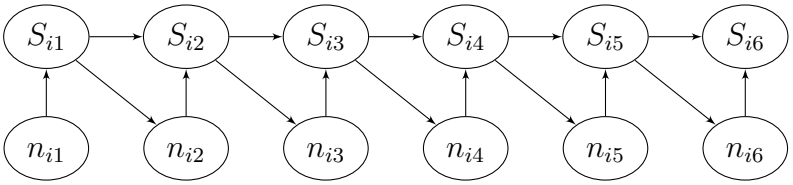


Figure 4.1: Network representation of a genotype at a marker with $A = 6$ allelic types.

cation happens according to a binomial distribution. In Proposition 2 we establish the formal correctness of the network specification.

Proposition 2. *The distributions of genotypes and partial sums satisfy the following relations*

$$S_{i1} = n_{i1},$$
$$n_{i1} \sim \text{bin}\,(2, q_1)\,,$$

and for $a \in \{2, \ldots, A\}$

$$S_{ia} = S_{i,a-1} + n_{ia},$$
$$n_{ia} \,|\, S_{i,a-1} \sim \text{bin}\left(2 - S_{i,a-1}, q_a / \textstyle\sum_{b=a}^{A} q_b\right). \tag{4.3}$$

Finally, we have the conditional independence relations

$$n_{ia} \perp\!\!\!\perp (n_{i1}, \ldots, n_{i,a-1}, S_{i1}, \ldots, S_{i,a-2}) \,|\, S_{i,a-1} \tag{4.4}$$
$$S_{ia} \perp\!\!\!\perp (n_{i1}, \ldots, n_{i,a-1}, S_{i1}, \ldots, S_{i,a-2}) \,|\, (S_{i,a-1}, n_{ia}).$$

Proof. The unnumbered relations follow directly from the definition of the quantities involved. We further have

$$\begin{aligned} p(n_{ia} \,|\, n_{i1}, \ldots, n_{i,a-1}) &= \frac{p(n_{i1}, \ldots, n_{i,a-1}, n_{ia})}{p(n_{i1}, \ldots, n_{i,a-1})} \\ &= \frac{\frac{2!}{(2-S_{i,a-1}-n_{ia})! \prod_{b=1}^{a} n_{ib}!} \left(\sum_{b=a+1}^{A} q_b \right)^{2-S_{i,a-1}-n_{ia}} \prod_{b=1}^{a} q_b^{n_{ib}}}{\frac{2!}{(2-S_{i,a-1})! \prod_{b=1}^{a-1} n_{ib}!} \left(\sum_{b=a}^{A} q_b \right)^{2-S_{i,a-1}} \prod_{b=1}^{a-1} q_b^{n_{ib}}} \\ &= \binom{2 - S_{i,a-1}}{n_{ia}} \left(1 - \frac{q_a}{\sum_{b=a}^{A} q_b} \right)^{2-S_{i,a-1}-n_{ia}} \left(\frac{q_a}{\sum_{b=a}^{A} q_b} \right)^{n_{ia}}. \end{aligned}$$

The conditional independence (4.4) follows from the fact that the conditional distribution of n_{ia} given $n_{i1}, \ldots, n_{i,a-1}$ only depends on the condition through $S_{i,a-1}$; inspection of the expression for the conditional distribution yields (4.3). □

A Bayesian network is defined on the Cartesian product of sample spaces, and in HUGIN we need to specify its conditional probability table for all combinations of the node and its parents, in particular also on configurations of the parents with probability 0. For combinations $n_{ia} + S_{i,a-1} > 2$, we give S_{ia} a uniform distribution on the states $\{0, 1, 2\}$, but note that the extension to parent configurations of probability 0 may be chosen arbitrarily.

4.2 Auxiliary variables for evaluation of the likelihood function

The likelihood function for peak heights can be found by marginalising over the unknown genotypes in the joint model of genotypes and peak heights, and we shall here rely on the expression (2.12) for the likelihood function,

$$L(\psi \,|\, H) = \prod_{m=1}^{M} \mathbb{E}\left\{ \prod_{a=1}^{A_m} f_\psi(z_a^m \,|\, \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m) \,\middle|\, H \right\}.$$

The inner expectation is seen to be a product over alleles within a marker, where each factor is a function of the variables $\boldsymbol{n}_a^m$ and $\boldsymbol{n}_{a+1}^m$, and so we can compute this expectation using auxiliary variables as described in Section 3.2. For each allele a, we add a binary auxiliary variable O_a with parents n_{ia} and $n_{i,a+1}$ for all unknown contributors i, except for O_A that is given only one parent n_{iA} per contributor.

In the notation of Section 3.2, O_a plays the role of Y^B, the corresponding parent set X_B is the set $(\boldsymbol{n}_a, \boldsymbol{n}_{a+1})$ of allele counts for alleles a and $a+1$, and the objective is to compute the expectation

$$\mathbb{E}\left\{\prod_{a=1}^{A_m} h_a(\boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m)\right\},$$

where $h_a(\boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m) = f_\psi(z_a^m \mid \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m)$.

Figure 4.2 shows the network for modelling one marker of a mixture with two contributors and six alleles. The structure displayed in Figure 4.2 can be seen as representing our model as coupled hidden Markov models.

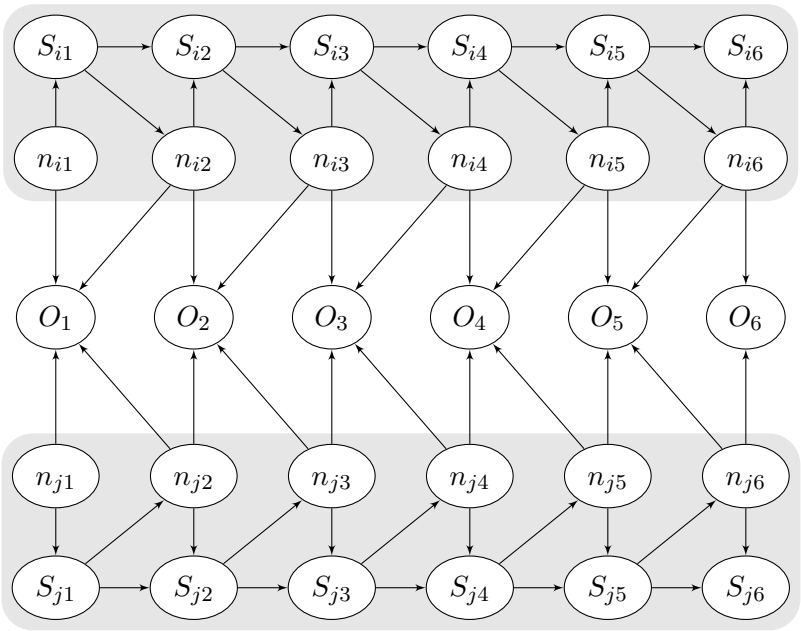


Figure 4.2: Bayesian network modelling the genotypes of 2 unknown contributors i and j for a marker with 6 possible allelic types.

Note that O_a and its two parents $n_{ia}, n_{i,a+1}$ for each unknown contributor i are necessarily contained in the same clique, implying that any valid junction tree will

contain cliques with an associated state space that is exponential in the number of unknown contributors. Unfortunately, as the moralised graph is not chordal – for instance $(S_{i1}, n_{i1}, n_{j2}, n_{i3}, S_{i2}, S_{i1})$ is a cycle – further edges need to be added, resulting in an additional increase in the size of the cliques. We shall return to this issue in Section 4.4.

The auxiliary variable O_a may be thought of as representing the observed peak height Z_a. To facilitate this interpretation, we define the distribution of O_a on the space $\{0, 1\}$ in a way that, for all alleles, the event $O_a = 1$ corresponds to the event $Z_a \geq C$, i.e. observing a peak above the threshold C. The height of the observed peak is taken into account through the distribution of O_a.

The conditional distribution of an auxiliary variable O_a given the allele counts is defined as follows. Using the conditional distribution (2.8) of the peak height Z_a given the allele counts, if a peak for allele a has been observed above the detection threshold C, i.e. $z_a \geq C$, the distribution of O_a is defined by

$$\mathrm{P}(O_a = 1 \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) = g_\psi(z_a \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1})/k_a^\psi, \tag{4.5}$$

with $k_a^\psi = \max_{\boldsymbol{n}_a, \boldsymbol{n}_{a+1}} g_\psi(z_a \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1})$, noting the dependence of the scaling factor k_a^ψ on ψ. For an undetected peak, i.e. $z_a^m = 0$, the distribution of O_a is defined by the conditional probability of not observing a peak,

$$\mathrm{P}(O_a = 0 \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) = G_\psi(C \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}). \tag{4.6}$$

Now Proposition 1 can readily be used to evaluate the likelihood for a given value of ψ by propagating likelihood evidence

$$\ell_a(O_a) = \begin{cases} k_a^\psi \mathbb{1}_{\{O_a=1\}}, & \text{if } z_a \geq C \\ \mathbb{1}_{\{O_a=0\}}, & \text{if } z_a < C. \end{cases} \tag{4.7}$$

4.3 Posterior distributions of genotypes

The inclusion of auxiliary variables may serve other purposes than merely as a device for calculating an expectation. For the interpretation of a DNA mixture, we are interested in the conditional distribution of the genotypes $\boldsymbol{n}$ for contributors given – possibly only a subset of – the observed peak heights.

By propagating likelihood evidence (4.7) for a set of alleles $\mathcal{A} \subseteq \{1, \ldots, A\}$, we obtain a representation of the conditional distribution of the full network given the relevant state of the auxiliary variables $(O_a, a \in \mathcal{A})$. We have defined the auxiliary variables so that for all alleles the event $O_a = 1$ corresponds to the event $Z_a \geq C$, i.e. that the peak at allele a is above the threshold C. Therefore, conditioning on auxiliary variables $(O_a, a \in \mathcal{A})$ yields the conditional distribution of the nodes in the network given the peak height information $\{z_a\}_{a \in \mathcal{A}}$, as formalised in the following:

Proposition 3. *Let X denote the set of network variables. For an arbitrary subset $\mathcal{A}$ of alleles we have*

$$p(x \mid \{z_a\}_{a \in \mathcal{A}}) = p\Big(x \,\Big|\, \bigcap_{\substack{a \in \mathcal{A} \\ z_a \geq C}} \{O_a = 1\} \bigcap_{\substack{a \in \mathcal{A} \\ z_a < C}} \{O_a = 0\}\Big). \tag{4.8}$$

Proof. This follows from the following argument:

$$
\begin{aligned}
p(x)\prod_{a\in\mathcal{A}}\ell_a(O_a) &\propto p\Big(x\,\Big|\bigcap_{\substack{a\in\mathcal{A}\\ z_a\geq C}}\{O_a=1\}\bigcap_{\substack{a\in\mathcal{A}\\ z_a<C}}\{O_a=0\}\Big)\\
&\propto p(x)\,\mathrm{P}\Big(\bigcap_{\substack{a\in\mathcal{A}\\ z_a\geq C}}\{O_a=1\}\bigcap_{\substack{a\in\mathcal{A}\\ z_a<C}}\{O_a=0\}\,\Big|\,x\Big)\\
&= p(x)\prod_{\substack{a\in\mathcal{A}\\ z_a\geq C}}\mathrm{P}(O_a=1\,|\,x)\prod_{\substack{a\in\mathcal{A}\\ z_a<C}}\mathrm{P}(O_a=0\,|\,x)\\
&= p(x)\prod_{\substack{a\in\mathcal{A}\\ z_a\geq C}}\mathrm{P}(O_a=1\,|\,\boldsymbol{n}_a,\boldsymbol{n}_{a+1})\prod_{\substack{a\in\mathcal{A}\\ z_a<C}}\mathrm{P}(O_a=0\,|\,\boldsymbol{n}_a,\boldsymbol{n}_{a+1})\\
&= p(x)\prod_{\substack{a\in\mathcal{A}\\ z_a\geq C}}\{g_\psi(z_a\,|\,\boldsymbol{n}_a,\boldsymbol{n}_{a+1})/k_a^\psi\}\prod_{\substack{a\in\mathcal{A}\\ z_a<C}}G_\psi(C\,|\,\boldsymbol{n}_a,\boldsymbol{n}_{a+1})\\
&\propto p(x)\prod_{a\in\mathcal{A}}f_\psi(\{z_a\}_{a\in\mathcal{A}}\,|\,x)\\
&\propto p(x\,|\,\{z_a\}_{a\in\mathcal{A}})
\end{aligned}
$$

as desired. □

4.3.1 Simulation under the model

As stated in Proposition 3, introducing evidence on the auxiliary variables O_a yields a representation of the posterior distribution of the genotypes of the unknown contributors. This in turn enables simulation of a full set of DNA profiles and corresponding peak heights, either marginally or conditionally on relevant subsets of the observed peak heights through standard Bayesian network methods (see Chapter 3). More generally, we have for any event B that

$$f_\psi(\{z_a\}_{a\in A},\boldsymbol{n}\,|\,B)=f_\psi(\{z_a\}_{a\in A}\,|\,\boldsymbol{n},B)p(\boldsymbol{n}\,|\,B).$$

If conditioning on the event B can be represented by propagation in our Bayesian network, for example if $B=\{Z_b=z_b,b\neq a\}$, we can readily simulate from $p(\boldsymbol{n}\,|\,B)$. Then, to sample the peak heights, we just further need a method for sampling from $f_\psi(\{z_a\}_{a\in A}\,|\,\boldsymbol{n},B)$.

This method of simulation can for example be used in a bootstrap analysis of the estimation uncertainty as in Section 6.1 below, or in a Monte-Carlo based fully Bayesian analysis as in Graversen and Lauritzen (2013).

4.4 Junction-tree representations and complexity considerations

The main concerns in computation by auxiliary variables are that the junction tree representation of the network may not fit in the physical memory, and that propagation and other network operations may take prohibitively long. Both of these issues are directly related to the total size of the network junction tree, as discussed in Chapter 3.

An additional concern lies in finding a good triangulation, as this can be both time- and memory-consuming; we eliminate this additional cost by specifying triangulations directly.

In the following we study the relation of the total sizes of junction tree representations used for DNA mixture analysis to the number A of possible alleles at a marker and the number k of unknown contributors. Known contributors, with their fixed genotypes, essentially do not affect the complexity and are therefore not included in the discussion.

4.4.1 Markov genotype representation

We shall consider three different triangulations of the network representation of the model based on a combination of the Markovian genotype representation in Section 4.1 and sets of auxiliary variables, such as those discussed in Section 4.2. For these triangulations, we investigate the behaviour of the total sizes of the corresponding junction trees. We restrict attention to mixture networks where any allele a – apart from the last allele A – can receive stutter from $a + 1$.

Any triangulation must necessarily have cliques that contain an auxiliary variable together with its parent set, as these sets are complete in the moralised graph. For all our junction trees we avoid adding additional variables to all such sets and simply combine any auxiliary variable with its parent set to form a clique. We can thus focus the discussion on triangulating the part of the moralised graph that does not involve auxiliary variables.

If we have N binary auxiliary variables per allele, their cliques and corresponding separators contribute to the total size of the junction tree by

$$TS_{\text{aux}} = 3N\left\{(A-1)3^{2k} + 3^k\right\},$$

since there are $N(A-1)$ cliques containing an auxiliary variable along with its $2k$ parents, and each is separated from the remaining junction tree by a separator containing the $2k$ parents. The N auxiliary variables for the last allele have only k parents.

Bearing Figure 4.1 in mind, the structure of the genotype networks requires *upper triangle* sets $\{S_{i,a-1}, S_{ia}, n_{ia}\}$ to be in a clique as they are complete sets. If allele $a-1$ receives stutter from a, then the *lower triangle* set $\{n_{i,a-1}, n_{ia}, S_{ia}\}$ is also complete in the moralised graph and must be contained in some clique.

The first triangulation method we shall consider, uses the simple idea of slicing the network into cliques

$$\{S_{ia}, S_{i,a+1}, n_{ia}, n_{i,a+1}\}_{i=1}^{k}$$

for $a = 1, \ldots, A-1$. The corresponding junction tree, which we shall refer to as the *slice tree*, is displayed in Figure 4.3.

We note that using the propagation algorithm on the slice tree is effectively equivalent to using the forward–backward algorithm on the hidden Markov chain with these cliques representing the hidden states. In addition to the cliques and separators arising from the auxiliary variables, the slice tree has $A-1$ cliques each consisting of $4k$ nodes, and $A-2$ separators between them, each consisting of $2k$

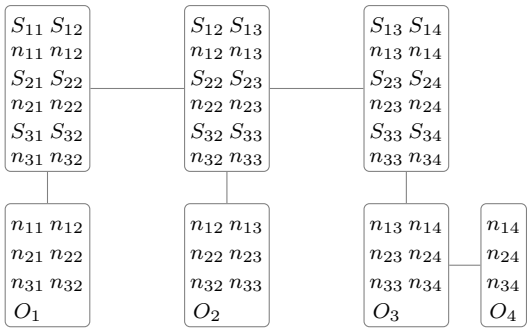


Figure 4.3: Slice junction tree for $k = 3$ contributors, $A = 4$ alleles, and $N = 1$ auxiliary variable per allele.

nodes. Thus the total size of the slice tree becomes

$$TS_{\text{slice}} = (A-1)3^{4k} + (A-2)3^{2k} + TS_{aux}.$$

However, we can improve on this triangulation by splitting each slice into two cliques as Figure 4.4 illustrates. The resulting *triangle tree* in Figure 4.5 has $2(A-1)$

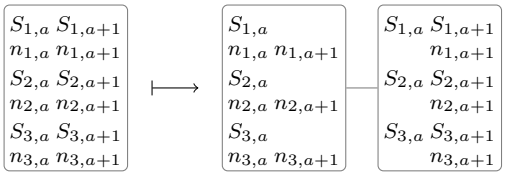


Figure 4.4: Splitting each slice into two cliques consisting of lower and upper for a reduction in total size.

cliques of each $3k$ nodes and $2(A-1)$ separators of each $2k$ nodes, and thus the total size

$$TS_{\text{triangle}} = 2(A-1)3^{3k} + \{2(A-1)-1\}3^{2k} + TS_{\text{aux}}$$

grows less quickly with the number of unknown contributors than the slice tree; see Figure 4.8.

In the case of only one unknown contributor, the total size of the triangle tree cannot be reduced. However, with more than one unknown contributor, each clique containing k upper triangles can be further split into k cliques as in Figure 4.6.

Note that the cliques containing k lower triangle sets cannot be split in a similar fashion. The resulting junction tree – the *split tree* – then has $A-1$ cliques of each

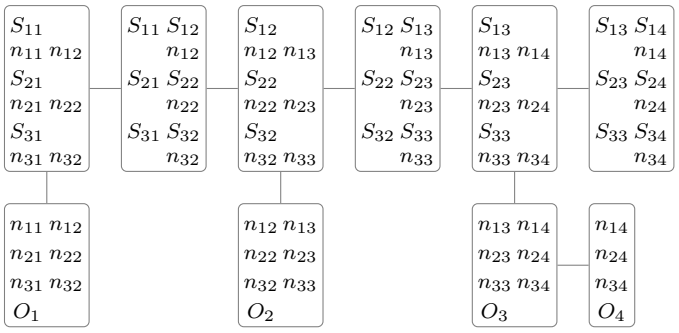


Figure 4.5: Triangle junction tree for $k = 3$ contributors, $A = 4$ alleles, and $N = 1$ auxiliary variable per allele.

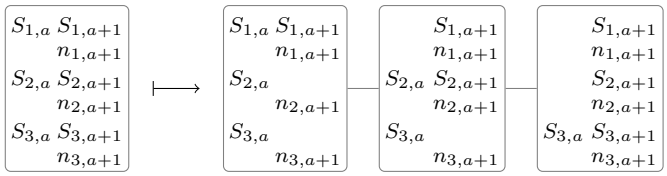


Figure 4.6: Splitting upper triangle cliques for a further reduction in total size.

$3k$ nodes, a further $k(A-1)$ of each $2k+1$ nodes, and $(k+1)(A-1)-1$ separators of $2k$ nodes between them. The total size of the tree is thus

$$TS_{\text{split}} = (A-1)3^{3k} + \{(4k+1)(A-1)-1\}3^{2k} + TS_{\text{aux}}.$$

A further slight and almost insignificant reduction of the total size of the split tree can be obtained by a small alteration in the cliques that cover nodes for the first two and last three alleles; the resulting tree is seen in Figure 4.7. The (slightly improved) split tree is the best junction tree we have been able to construct. We have investigated junction trees found by the algorithm for minimizing total clique size (excluding separators) as implemented in HUGIN, but none have smaller total size than the split tree.

The exponential growth of the total size is illustrated in Figure 4.8 for each of the three types of junction tree. Our numerical examples all include $N = 3$ auxiliary variables for each allele to reflect the size of the networks used in the R package DNAmixtures (Graversen, 2014). The number N of auxiliary variables makes little difference to the total size, as this in all cases grows linearly with N.

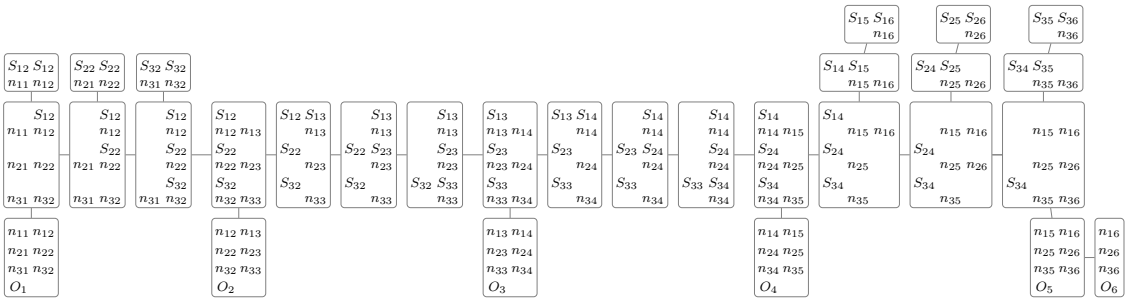


Figure 4.7: The best of our junction trees, the split tree, here for a DNA mixture network with $k = 3$, $A = 6$, and $N = 1$.

An advantage of the split tree is that it can be generated by an *elimination sequence*; such a sequence of nodes determines a set of edges to add to the moralised graph so that a triangulated graph is obtained. For each node v_i in the elimination sequence, edges are added to $\mathcal{G}$ so that in the subgraph of $\mathcal{G}$ obtained by eliminating nodes $\{v_1, \ldots, v_{i-1}\}$ and their adjacent edges, the set of neighbours of v_i becomes a complete set. When the last node of the elimination sequence is reached, the graph $\mathcal{G}$ is triangulated and a junction tree of the cliques may be constructed.

The elimination sequence for the split tree first eliminates all the auxiliary variables and then proceeds through the network nodes as

$$\boldsymbol{S}_A, \boldsymbol{S}_{A-1}, \boldsymbol{S}_1, \boldsymbol{n}_1, \{\boldsymbol{n}_a, \boldsymbol{S}_a\}_{a=2}^{A-2}, \boldsymbol{n}_{A-1}, \boldsymbol{n}_A \tag{4.9}$$

where $\boldsymbol{S}_a$ denotes $\{S_{ia}\}_{i=1}^{k}$ etc.

All of the triangulations found by the methods implemented in HUGIN can be constructed by an elimination order, and this is also the format in which triangulations can be specified by the user.

Compression

The network representations constructed for the genotypes have a large number of configurations that are impossible, for example due to the constraint that $\sum_a n_{ia} = 2$ for all i. In HUGIN there is a facility to *compress* the domain, such that only configurations of clique and separator states with non-zero probability are stored,

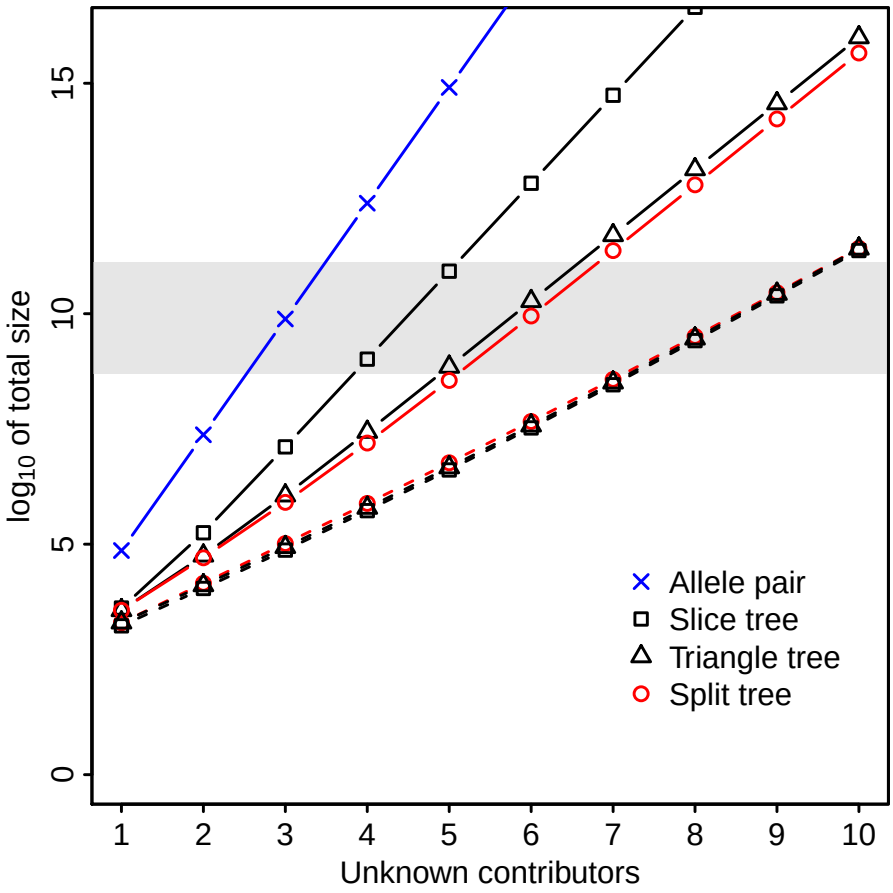


Figure 4.8: Total sizes of junction trees as a function of the number k of unknown contributors, in the case of $A = 25$ allelic types and $N = 3$ auxiliary variables per allele. Solid lines are uncompressed sizes and dashed lines compressed sizes. The horizontal band indicates total sizes ranging from 2GB to 512GB assuming numbers are represented in single precision.

thus reducing the effective size of the junction tree. There is a slight cost in terms of bookkeeping, but for our purposes this cost is negligible.

As is apparent from Figure 4.8, the exponential growth pattern prevails for the compressed domains. Note that after compression, all three junction trees are approximately of the same size. Also, the reduction of total size obtained by compression is itself growing exponentially; ignoring any slight reduction in total size from compressing states with probability zero in the cliques with auxiliary variables, the total size for the compressed slice tree is

$$TS_{\text{compr.slice}} = (A-3)10^k + \{3N(A-1) + A\}\,6^k + 3N3^k.$$

To make a compression, one single propagation has to be performed and therefore the uncompressed networks set the limit for computational feasibility. When numbers are represented in single precision of each four bytes, the horizontal band in Figure 4.8 represents a range of capacities from 2GB to 512 GB of memory. The band indicates that using the split junction tree should enable computation for up to $k = 6$ unknown contributors, whereas using the slice tree restricts computation to around $k = 4$.

There is a simple way of compressing the slice tree in that there are at most 10 possible configurations of the states in each of $\{S_{ia}, S_{i,a+1}, n_{ia}, n_{i,a+1}\}$. So if the state space is defined by these from the outset, it would in principle be possible to handle up to $k = 9$ unknown contributors, as then the compressed network would determine the maximal capacity; however, the general flexibility of the network representation would be reduced.

According to Steele and Balding (2014), four of the six programs currently available can handle up to 3-4 unprofiled contributors in the evaluation of the WoE, while DNAmixtures and the commercial software TrueAllele are able to handle up to 6; however, the computational resources needed to achieve this are not stated.

A suitable benchmark for empirical performance analyses is difficult to establish, even just for comparing run-times within one piece of software, as it depends heavily on the problem and the specific implementation. However, we note that all analyses in this thesis have been made using DNAmixtures on a standard desktop computer. Further, we note that our practical experience complies with the above theoretical analyses in that maximisation of the likelihood in models with up to five unknown contributors is amply feasible using the desktop computer, whereas the model with six unknown contributors was performed on a computer with 300GB RAM – mainly due to the size of the networks before compression.

4.4.2 Alternative genotype representations

The network representation of the model illustrated in Figure 4.2 consists of the genotype networks for unknown contributors tied together in the sense that they each contain parent nodes for the various auxiliary variables. Clearly, the network representing the genotype of an unknown contributor could be substituted for a different representation than the one suggested here and connected to the auxiliary variables in an appropriate way.

We briefly consider two alternative representations of a genotype. The associated total size of the junction trees are in both cases exponential in the number k of unknown contributors; see Graversen and Lauritzen (2014) for further details.

Representation by a pair of alleles. More commonly, a genotype is represented directly as an unordered pair of alleles; this representation has for example been used in Cowell et al. (2011). Using a single node to represent the pair of alleles for an unknown contributor, the parent sets of the auxiliary variables contain k genotype nodes each with $A(A+1)/2$ possible states for a marker with A possible alleles. The total size is in this representation polynomial rather than linear in the number A of alleles and the exponential growth rate in k increases with A; Figure 4.8 gives the total size for $A = 25$ alleles.

Representation by single alleles. Another possibility, used for example in Dawid et al. (2002) and Mortera et al. (2003), is to model the genotype at the level of the single alleles. Representing an allele by the vector of allele counts, this has a multinomial distribution with appropriate allele frequencies and a total of one allele to allocate. Thus, the allele can be represented by the Markov structure in Figure 4.1 used for a genotype, just that each node n_{ia} or S_{ia} has state space $\{0, 1\}$ rather than $\{0, 1, 2\}$.

The representation by single alleles comes with a cost, in that two networks are needed per unknown contributor. But, although less efficient the representation by single alleles may be preferable for specifying more complex models for unknown genotypes, for instance in an analysis along the lines of Green and Mortera (2009).

4.5 Extensions and modifications

The general flexibility of the Bayesian network representation can be exploited in many ways. Below we give a few examples of extensions and modification to the DNA mixture model that accommodate some commonly desirable features.

4.5.1 Representation of multiple mixtures

In the joint model for mixtures, Model 2.3, the peak heights for different mixtures are assumed conditionally independent given the genotypes of the set contributors shared between mixtures, and independent of the genotypes of non-shared contributors. The Bayesian network representation of the model is thus readily extended to the scenario of multiple mixtures by including a set of auxiliary variables $(O^m_{e1}, \ldots, O^m_{eA_m})$ for each EPG e that includes marker m.

4.5.2 Representation of the qualitative model

The qualitative model presented in Section 2.5 can also be represented by the use of auxiliary variables. We introduce binary variables D_a that are given the same parents as for O_a and with conditional probability

$$\mathrm{P}(D_a = 1 \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) = \mathrm{P}(Z_a \leq C \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}). \tag{4.10}$$

so that the probability of the event $D_a = 1$ equals the probability of a peak falling below the peak-height threshold C. The computation of the likelihood function as well as of the posterior distribution of genotypes given observations is now completely

analogous to the methodology developed in Sections 4.2 and 4.3, simply using the auxiliary variables D_a in place of O_a.

4.5.3 Amelogenin

As described in Section 1.1, a person has either two copies of the X chromosome (female) or a single copy of each of the X and Y chromosomes (male). Letting p_{XX} and p_{XY} denote the frequencies of females resp. males in the population, the possible states of the allele counts for Amelogenin are $(n_{iX}, n_{iY}) \in \{(2,0),(1,1)\}$ with probabilities

$$\begin{aligned} \mathrm{P}((n_{iX}, n_{iY}) = (2,0)) &= p_{XX} \\ \mathrm{P}((n_{iX}, n_{iY}) = (1,1)) &= p_{XY} = 1 - p_{XX}. \end{aligned}$$

The number $n_{iX} \in \{1,2\}$ of alleles X completely determines the sex and has a distribution

$$\mathrm{P}(n_{iX} = 2) = 1 - \mathrm{P}(n_{iX} = 1) = p_{XX}.$$

The number of alleles of Y, $n_{iY} = 2 - n_{iX} = 2 - S_{iX}$ is trivially independent of n_{iX} conditionally on S_{iX}. Conditionally on the number S_{iX} of alleles X, the number n_{iY} follows a binomial distribution with count $2 - n_{iX}$ and probability $1 = p_{XX}/p_{XX}$ as for the Markov genotype representation. Further $S_{iY} = 2$ is deterministic and therefore trivially independent of n_{iX} given the parent nodes (S_{iX}, n_{iY}). Thus Amelogenin is representable by the network shown in Figure 4.9, which has the exact same structure as other markers. This has the immediate advantage that the methodology for further computations in the model readily apply to Amelogenin.

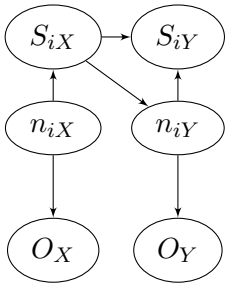


Figure 4.9: Markov representation of a genotype at the marker Amelogenin

4.5.4 Silent alleles

As discussed in Section 2.2.3, there can be a need for different ways of modelling dropout in more extreme cases where it cannot be explained adequately by the general variability of the gamma distribution.

The possibility that the unknown contributors has a silent allele can be incorporated in the model by adding an extra allelic type 0 (silent) to the model with some suitable allele frequency q_0 and re-normalising all other allele frequencies to sum to $1-q_0$. This corresponds to a modification of the basic model where the allele frequencies $q_a, a > 0$ are interpreted as the relative frequencies of non-silent alleles, so the probability that a random allele is of type a is equal to $(1 - q_0)q_a$ if $a > 0$.

We may think of the modified model for a genotype as a modification of an existing network representation covering alleles $1, \ldots, A$, to which we add $n_{i0} \sim \text{Bin}(2, q_0)$ and $S_{i0} = n_{i0}$; the extension is illustrated in Figure 4.10, where the grey shaded area marks the original genotype representation. The silent allele is special in the sense that no peak can be observed, and therefore no observational node O_0 is included for this allele.

Conditionally on the number S_{i0} of silent alleles allocated, the allele counts $(n_{i1}, \ldots, n_{iA})$ for the non-silent alleles are multinomially distributed with a total of $2 - S_{i0}$ alleles and frequencies

$$(q_1(1 - q_0), \ldots, q_A(1 - q_0))/(1 - q_0) = (q_1, \ldots, q_A).$$

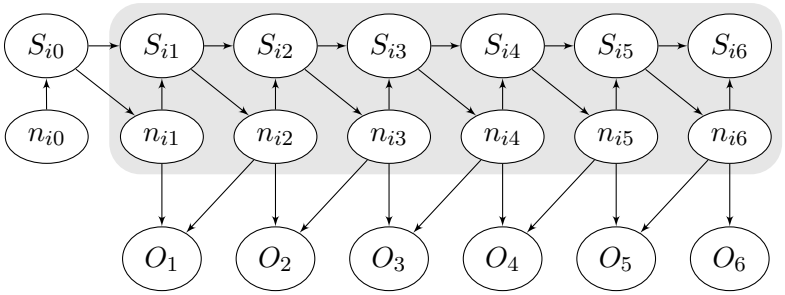


Figure 4.10: Modified network allowing for silent alleles. The nodes in the shaded area are those in the original genotype representation.

Thus, all we need is to change in the existing network is the total number of alleles to allocate, which is now $2 - S_{i0}$ rather than 2. Now $S_{i1} = n_{i1} + S_{i0}$ and $n_{i1} \mid S_{i0} \sim \text{Bin}(2 - S_{i0}, q_1)$, where q_0 is the probability of an allele being silent and q_1 indicates the original database allele frequency for allele 1. Conditionally on S_{i1}, the remaining allocation of alleles is independent of the allocation of silent alleles, and therefore the rest of the existing network remains unchanged.

Chapter 5

Statistical interpretation of a DNA mixture

Chapters 2-4 saw the development of a statistical model for DNA mixtures as well as the development of computational methodology for implementing the model, ensuring that the model can be of practical use.

We now change the focus to the development of methodology relating to the further statistical inference and discuss how the model can be useful for DNA mixture analysis.

Example 5.1 (DNA mixtures MC15 and MC18). For illustration of our methodology, we use the two DNA mixtures MC15 and MC18 introduced in section 1.3. The analysis here extends the example analyses of Cowell et al. (2015) and Graversen and Lauritzen (2014). For the sake of consistency with their analyses we use the US-Caucasian allele frequencies (Butler et al., 2003) throughout, with no further adjustments.

In the following we let the discussion revolve around the two hypotheses

$$H_p(4) : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, U_1$$
$$H_d(4) : K_1 \,\&\, K_2 \,\&\, U_1 \,\&\, U_2,$$

corresponding to a scenario where K_3 is the defendant and the profiled individuals K_1 and K_2 are both included as potential contributors. For the joint analysis of MC15 and MC18, these hypotheses specify their joint set of contributors, thereby following closely the setting of Model 2.3 for joint models of mixtures.

More generally, we let $H_p(k)$ denote the prosecution hypothesis involving a total of k contributors, of which K_1, K_2, and K_3 are profiled. Similarly, let $H_d(k)$ denote the defence hypothesis involving k contributors, of which contributors K_1 and K_2 are profiled. For ease of notation, unknown contributors are denoted by $U_1, \ldots, U_N$ for any hypothesis with N unknown contributors, but it should be emphasised that these labels are merely placeholders and that neither genotypes nor the proportions of DNA from the associated contributors are assumed to be the same across hypotheses.

•

5.1 Maximum-likelihood estimation

The statistical model for peak heights involves for a given hypothesis a set of model parameters $\psi = (\rho, \eta, \xi, \phi)$, which need to be specified in order for the model to be operative.

A preliminary study of estimation methods was presented in Graversen and Lauritzen (2013) for the gamma model for relative peak heights (Cowell et al., 2007a) in the setting of two contributors. The study indicated that it is possible to estimate both the mixture proportions and a parameter similar to our ρ based solely on the information in a single DNA mixture; indeed, the likelihood was seen to be

highly peaked. Here we extend the discussion to the setting of the more elaborate model with artefacts, allowing also more contributors than just two.

The vector ϕ of mixture proportions is an important part of the explanation of the trace and would generally be case specific. An exception to this would be in modelling replicate analyses of the same sample, as here it would be of interest to constrain the mixture proportions to be equal across the replicates. In theory, ρ is proportional to the amount of DNA in the sample and therefore depends on the case at hand. The stutter proportion ξ and the scale parameter η would be expected to depend on the laboratory procedure and settings, but to be common across cases analysed under the same circumstances.

In the following, we consider the number of contributors fixed and known and discuss the estimation of $\psi = (\rho, \eta, \xi, \phi)$ by the method of maximum likelihood. Interchangeably, we use also the parametrisation (μ, σ, ξ, ϕ) with $\mu = \rho\eta$ and $\sigma = 1/\sqrt{\rho}$ because of its more direct interpretability; we recall from Chapter 2 that μ is the mean and σ is the coefficient of variation for a peak height in the case of a single heterozygous contributor.

The parameter space

For each DNA mixture included in the model we let $\rho > 0$, $\eta > 0$, and $0 \leq \xi < 1$. The parameter space for mixture proportions is the unit k-simplex, so $0 \leq \phi_i \leq 1, i = 1, \ldots, k$ with the constraint $\sum_{i=1}^{k} \phi_i = 1$. Note that by allowing $\phi_i = 0$, we cover the possibility that a proposed individual has not contributed any DNA to the mixture. Also, by allowing $\phi_i = 1$ it is possible to model DNA samples originating from just a single contributor.

To ensure the identifiability of the parameters, we further order any unknown contributors according to non-increasing contributions to the mixture, such that if $i < j$ for two unknown contributors i and j, then $\phi_i \geq \phi_j$.

In case of the joint analysis of multiple mixtures, the order restriction of mixture proportions for unknown contributors is only imposed for the first mixture (see Section 2.4). This allows the order of the contributions to be different across the mixtures, but may not fully guarantee identifiability of the parameters – two mixture proportions may be equal under the first mixture, but not under the subsequent mixture(s). Introducing further order restrictions on the mixture proportions would not cause any limitations for the further inference, but the choice of appropriate restrictions depends on the hypothesis under consideration.

Maximisation

For the maximisation of the likelihood we use the numerical method `solnp` in the Rsolnp package for R (Ghalanos and Theussl, 2012). This implements a nonlinear augmented Lagrange multiplier method suggested by Ye (1987). For any real-valued nonlinear smooth function $f : \mathbb{R}^n \to \mathbb{R}$, the method solves the problem of minimising $f(x)$ under the constraints

$$l \leq x \leq u, \text{ with } l < u$$
$$l_h \leq h(x) \leq u_h, \text{ with } l_h < u_h$$
$$g(x) = \mathbf{0}.$$

Here $g : \mathbb{R}^n \to \mathbb{R}^{m_1}$ and $h : \mathbb{R}^n \to \mathbb{R}^{m_2}$ are vector-valued nonlinear smooth functions.

The flexibility in imposing constraints on the parameters, makes `solnp` highly suitable for our somewhat exploratory purpose. The main aim is to investigate whether the model is generally useful and whether it is possible to estimate the parameters in practice. We use the default settings of `solnp` as these seem to work satisfactorily and produce sensible and stable results. In the future, it may be worth adapting the settings to the specific problem as well as experimenting with other optimisation methods and parameterisations of the likelihood function.

Example 5.2 (Estimation under four-person hypotheses for MC15). We firstly consider models for the mixture MC15. Table 5.1 shows the maximum-likelihood estimates under all the possible hypotheses involving a total of four contributors, among which are a number of the individuals K_1, K_2, and K_3.

Table 5.1: Maximum-likelihood estimates under all four-person hypotheses for explaining MC15.

Known contr.	ρ	η	μ	σ	ξ	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}	ϕ_{U_1}	ϕ_{U_2}	ϕ_{U_3}	ϕ_{U_4}	$\log L$
	24.7	37.0	915	0.201	0.066				0.795	0.069	0.068	0.068	-330.0
K_1	24.7	37.0	915	0.201	0.066	0.795			0.068	0.068	0.068		-297.9
K_2	25.5	35.8	914	0.198	0.072		0.039		0.798	0.081	0.081		-329.9
K_3	31.6	28.9	914	0.178	0.068			0.120	0.817	0.031	0.031		-308.8
K_1 & K_2	25.5	35.8	914	0.198	0.072	0.798	0.039		0.081	0.081			-297.8
K_1 & K_3	31.6	28.9	914	0.178	0.068	0.817		0.120	0.031	0.031			-276.8
K_2 & K_3	34.2	26.7	913	0.171	0.074		0.047	0.124	0.821	0.008			-303.9
K_1 & K_2 & K_3	34.2	26.7	913	0.171	0.074	0.821	0.047	0.124	0.008				-271.8

The estimates of ρ, η, μ, σ, and ξ are similar across the different hypotheses, as are the estimates of the mixture proportions for known contributors. The mixture proportions for unknown contributors cannot be directly compared, as the interpretation of unknown contributors depends on the hypothesis under investigation. However, we do see that there is a well determined major contribution of about 80% coming from either K_1 or, if that individual is not included, from the unknown contributor U_1. Gill et al. (2008) estimated the mixture proportions under their prosecution hypothesis $H_p(4) : K_1$ & K_2 & K_3 & U_1 and got an estimate of $(\phi_{K_1}, \phi_{K_2}, \phi_{K_3}, \phi_{U_1}) = (0.80, 0.06, 0.12, 0.04)$ similar to the estimates in Table 5.1. We note from the maximised likelihood that the hypothesis $H_p(4)$ offers the best explanation of the sample under the assumption that it contains at most DNA from four individuals. •

Example 5.3 (Estimation under four-person hypotheses for MC18). Turning now to models for mixture MC18, Table 5.2 gives the maximum likelihood estimates under for all four-person hypotheses as in Example 5.2.

Table 5.2: Maximum-likelihood estimates under all four-person hypotheses for explaining MC18.

Known contr.	ρ	η	μ	σ	ξ	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}	ϕ_{U_1}	ϕ_{U_2}	ϕ_{U_3}	ϕ_{U_4}	$\log L$
	21.5	49.1	1056	0.216	0.074				0.678	0.107	0.107	0.107	-371.1
K_1	21.7	48.6	1056	0.215	0.075	0.679			0.107	0.107	0.107		-339.1
K_2	33.8	31.2	1056	0.172	0.085		0.096		0.698	0.193	0.013		-362.2
K_3	28.8	36.7	1058	0.186	0.076			0.189	0.707	0.052	0.052		-347.2
K_1 & K_2	33.8	31.2	1056	0.172	0.085	0.698	0.096		0.193	0.013			-330.1
K_1 & K_3	28.8	36.7	1058	0.186	0.076	0.707		0.189	0.052	0.052			-315.1
K_2 & K_3	36.2	29.1	1056	0.166	0.085		0.091	0.194	0.706	0.009			-331.6
K_1 & K_2 & K_3	36.2	29.1	1056	0.166	0.085	0.706	0.091	0.194	0.009				-299.5

Also for this DNA mixture the estimates are similar across the varying hypotheses. A major contribution of around 70% is consistently determined and attributed to either K_1 or, when K_1 is not included, the major unknown contributor. The estimated mixture proportions correspond well to those stated in Gill et al. (2008), $(\phi_{K_1}, \phi_{K_2}, \phi_{K_3}, \phi_U) = (0.67, 0.11, 0.19, 0.04)$. As for MC15, $H_p(4)$ is the best of the offered explanations of the mixture MC18. •

Estimated mixture proportions

Tables 5.1 and 5.2 highlight an important issue with maximisation in our model: for many of the models the estimated mixture proportions are equal for the unknown contributors.

The peak height distribution for fixed genotypes depends on the genotypes only through $\sum_i \phi_i n_{ia}$. Thus, if two unknown individuals i, j have contributed the same amount of DNA so that $\phi_i = \phi_j$, then their total of four alleles may be allocated arbitrarily between the two individuals without changing the value of the density. Further, under the standard model for genotypes, the joint distribution of genotypes of the unknown contributors is symmetric in the genotypes.

This creates a symmetry of the likelihood function around $\phi_i = \phi_j$ for unknown contributors i and j. Symmetry in unknown contributors is eliminated by ordering

the unknown contributors after their mixture proportion. However, this does not eliminate the possibility of local or global maximum points on the boundary, which then result in equal mixture proportions.

Of course, as a function of many parameters, the likelihood function is difficult to inspect graphically. One possibility of assessing whether a suggested maximum is global is to use the fact that the global maximum for the likelihood will by necessity also be the global maximum for the profile likelihood for some of the parameters. Since the profile likelihood is itself found by numerical maximisation, the risk of encountering local modes evidently persists.

Example 5.4 (Checking for local modes). Figure 5.1 shows the profile log likelihood for $\phi_{U_1} - \phi_{U_2}$ in the model $H_d(4) : K_1 \,\&\, K_2 \,\&\, U_1 \,\&\, U_2$. The plot confirms that there is a global maximum for $\phi_{U_1} = \phi_{U_2}$; a detailed inspection of the profile likelihood outside the displayed range confirms that it continues to decrease as $\phi_{U_1} - \phi_{U_2}$ increases.

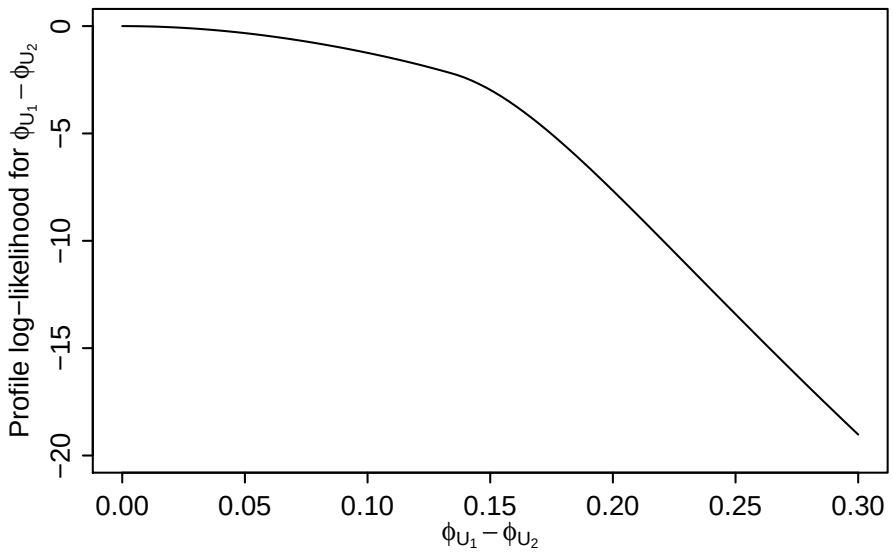


Figure 5.1: Profile log likelihood for $\phi_{U_1} - \phi_{U_2}$ under $H_d(4)$ for MC15.

•

Due to the possibility of encountering a local optimum, it is highly recommended to start the optimisation from multiple points and choose the best of the found optima.

Starting values

Although a convenient starting point, generally it is best to avoid starting the maximisation from even mixture proportions as this increases the risk of getting stuck in an optimum that is due to symmetry. However, such starting points may assist in picking up well determined contributions, which can then be used as a basis for constructing better starting values. As the number of contributors increases, it naturally gets increasingly difficult to choose very distinct proportions. Further, it can be important to establish the correct order of the mixture proportions; the order between unknown contributors is established, but as the number of known contributors increases so does the number of possible orders of all the contributors.

In a model with known genotypes and where no threshold is used, we can easily maximise the profile likelihood for η to find that

$$\hat{\eta} = \sum_{m,a} Z_a^m / (2M\hat{\rho}). \tag{5.1}$$

In this simple model,

$$\hat{\mu} = \hat{\rho}\hat{\eta} = \hat{\rho}\frac{\sum_{m,a} Z_a^m}{2M\hat{\rho}} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2}\sum_a Z_a^m. \tag{5.2}$$

Thus, $\hat{\mu}$ is an average peak height and depends only on the observed peak heights. In particular, as it does not depend on the genotypes, the maximum will be the same also when integrating over the genotypes of unknown contributors. Once some of the peak heights fall below threshold, the MLE for μ is more complicated and is dependent on both genotypes and other parameter estimates. However, the average peak height could provide a good starting value for the maximisation.

From Example 5.2 we note that the estimates $\hat{\mu}$ are highly similar across the various hypotheses, which would be expected if (5.2) is roughly true, as this expression does not depend on the hypothesis. For MC15 the average peak height calculated by (5.2) is 899, which is a bit lower than the estimates for μ around 914. The same holds for MC18, where the average peak height is 1049 and the estimates in Example 5.2 are 1056 or 1058.

We hope to address in the future the problem of automatically finding good starting values for the parameters, obviating the need for user-specified starting values as currently required in the R-package DNAmixtures.

Example 5.5 (Estimation in joint models for MC15 and MC18). Assuming that the two mixtures share a total of four contributors, we follow Examples 5.2 and 5.3 and estimate the parameters under various joint models for MC15 and MC18. We note from Table 5.3 that equal proportions among unknown contributors here only occurs for two models, $K_3 \,\&\, U_1 \,\&\, U_2 \,\&\, U_3$ and $K_1 \,\&\, K_3 \,\&\, U_1 \,\&\, U_2$.

Table 5.3: Maximum likelihood estimation in joint models for MC15 and MC18 involving a total of four contributors. Although stated twice, $\log L$ denotes the maximal value of the likelihood function for the joint model.

(a) Estimates for the parameters corresponding to MC15.

Known contr.	ρ	η	μ	σ	ξ	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}	ϕ_{U_1}	ϕ_{U_2}	ϕ_{U_3}	ϕ_{U_4}	$\log L$
	31.2	29.3	914	0.179	0.071				0.803	0.111	0.063	0.024	-649.5
K_1	31.2	29.3	914	0.179	0.071	0.803			0.111	0.063	0.023		-617.4
K_2	33.1	27.6	913	0.174	0.074		0.049		0.819	0.123	0.010		-635.9
K_3	33.4	27.4	915	0.173	0.068			0.119	0.815	0.033	0.033		-621.2
K_1 & K_2	33.1	27.6	913	0.174	0.074	0.819	0.049		0.123	0.010			-603.8
K_1 & K_3	33.6	27.2	915	0.173	0.068	0.815		0.119	0.033	0.033			-589.2
K_2 & K_3	34.2	26.7	913	0.171	0.074		0.047	0.124	0.820	0.008			-603.4
K_1 & K_2 & K_3	34.2	26.7	913	0.171	0.074	0.820	0.047	0.124	0.008				-571.3

(b) Estimates for the parameters corresponding to MC18.

Known contr.	ρ	η	μ	σ	ξ	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}	ϕ_{U_1}	ϕ_{U_2}	ϕ_{U_3}	ϕ_{U_4}	$\log L$
	28.9	36.6	1057	0.186	0.080				0.686	0.177	0.103	0.035	-649.5
K_1	28.9	36.6	1057	0.186	0.080	0.686			0.177	0.103	0.034		-617.4
K_2	35.4	29.9	1056	0.168	0.085		0.092		0.703	0.194	0.011		-635.9
K_3	30.2	35.0	1058	0.182	0.075			0.188	0.704	0.054	0.054		-621.2
K_1 & K_2	35.3	29.9	1056	0.168	0.085	0.703	0.092		0.194	0.011			-603.8
K_1 & K_3	30.2	35.0	1058	0.182	0.075	0.704		0.188	0.054	0.054			-589.2
K_2 & K_3	36.2	29.1	1056	0.166	0.085		0.091	0.194	0.706	0.010			-603.4
K_1 & K_2 & K_3	36.3	29.1	1056	0.166	0.085	0.706	0.091	0.194	0.010				-571.3

The estimates are similar to those obtained by analysing MC15 and MC18 separately, and we note that $H_p(4) : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, U_1$ remains the best explanation for the combination of the two mixtures under the assumption that a total of at most four people have contributed to the samples. •

5.2 Weight of Evidence

An important practical use of the likelihood function is for evaluation of the weight of evidence as introduced in Section 1.2.1. We evaluate the WoE as the $\log_{10}$ of the ratio of maximised likelihoods. Thus, for each of the competing hypotheses H we

compute

$$\hat{L}(H) = \sup_{\rho,\xi,\phi,\eta} L(\rho, \xi, \phi, \eta \,|\, \boldsymbol{z}, H)$$

which corresponds to using maximum likelihood estimates for the unknown parameters as estimated under the hypothesis H.

5.2.1 Quantitative peak information

Example 5.6 (Weight of Evidence, MC15). Using the mixture MC15, we investigate the weight of evidence against each of the three known individuals K_1, K_2, and K_3. For this, we consider the four-contributor hypotheses for MC15 and the corresponding parameter estimates in Table 5.1.

For each individual K of interest we compare two hypotheses; one in which the individual K is included, and one in which K is replaced by an unknown contributor. For the three remaining contributors, we then consider four explanations formed by including none, one, or both of the two other known profiles and as many unknown profiles as needed. The results are seen in Table 5.4.

Table 5.4: Weight of Evidence (WoE) against each of the three profiled individuals K_1, K_2, and K_3, using MC15 and a total of four contributors.

Defendant K	Other known contr.	WoE(K)
K_1		13.9
	K_2	13.9
	K_3	13.9
	K_2, K_3	13.9
K_3		9.2
	K_1	9.2
	K_2	11.3
	K_1, K_2	11.3
K_2		0.1
	K_1	0.1
	K_3	2.2
	K_1, K_3	2.2

The evidence against K_1 is strong and attains the upper bound $\text{WoE}_{\max}(K_1) = 13.9$ bans, corresponding to $\log_{10}$ of the inverse match probability for the profile of K_1. In particular, this means that there is no evidential loss and that the mixture would fully identify the contributor K_1 in a mixture deconvolution, c.f. (1.8).

The evidence against K_3 is also strong and does not depend on whether K_1 is assumed present or not – perhaps not surprising since the sample fully identifies one contributor as K_1. In contrast, the WoE does depend on whether K_2 is included in the hypothesis: If K_2 is not included, then the WoE against K_3 is 9.2, but by including K_2 the WoE is increased to 11.3 bans. This stronger WoE was reported in the analysis of Cowell et al. (2015). The evidence against K_3 is weaker than the upper bound of 14.5 bans.

Similarly to above, the WoE against K_2 does not depend on the inclusion of K_1, whereas inclusion of K_3 increases the WoE against K_2 from 0.1 bans to 2.2 bans. This is very weak evidence compared to the upper bound of 14.2 for a perfect trace. Assuming instead that MC15 is a mixture of at most three contributors, interestingly the WoE of -0.08 against K_2 now leans slightly toward K_2 not having contributed to the sample, but if K_3 is included then the WoE increases to 2.4 bans, leaning again toward K_2 having contributed to the sample. We omit the details on three-person hypotheses.

This example illustrates the fact that the evidence against a person depends on the specific hypotheses under consideration; there is no general rule for the effect of assuming the presence of DNA from one or more specific individuals. It may therefore be sensible to report more than a single evidential value. •

A combined analysis of multiple DNA samples can be highly informative, but poses new challenges in that it requires a suitable model for the shared set of contributors between the mixtures. The origin of the stains may provide reasons for a particular assumption about the unknown contributors: For replicate analyses of

the same stain or samples of same origin it could be reasonable to assume that all unknown contributors are shared, whereas for stains obtained far apart it may be more reasonable to assume distinct (independent) sets of unknown contributors.

Further, there is a wide range of possible hypotheses to compare in assessing the evidence against a specific individual. For instance the defence hypothesis could replace the defendant by a single shared unknown contributor or – at the other extreme – it could replace the defendant by one independent unknown contributor per mixture.

Example 5.7 (Combined WoE against K_3, MC15 and MC18). We now discuss two possible combined analyses of mixtures MC15 and MC18 for evaluating the weight of evidence against the defendant K_3. In both analyses, we use the two hypotheses $H_p(4) : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, U_1$ and $H_d(4) : K_1 \,\&\, K_2 \,\&\, U_1 \,\&\, U_2$ for the separate analyses of MC15 and MC18 to form two joint hypotheses that describe the joint set of contributors to the two mixtures.

Replacing K_3 by a shared unknown. Firstly, we form the joint prosecution hypothesis $H_p(4) : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, U_1$, simply letting U_1 be a potential contributor to both mixtures. The joint defence hypothesis $H_d(4) : K_1 \,\&\, K_2 \,\&\, U_1 \,\&\, U_2$ explains the mixtures by the presence of two shared unknown contributors, thus in essence replacing K_3 by a single unknown contributor who is present in both mixtures. The WoE is in this case 14.1 bans, which is close to the upper limit of 14.5 bans for a perfect match in a single-source sample.

Independent sets of unknown contributors. Another possibility is to form a joint prosecution hypothesis H_p' by combining $H_p(4) : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, U_1$ for MC15 with $\tilde{H}_p(4) : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, \tilde{U}_1$ for MC18 under the assumption that U_1 and $\tilde{U}_1$ are two distinct individuals that have each contributed to one of the mixtures only.

Similarly, we may form a joint defence hypothesis H'_d, under which K_1 and K_2 have contributed to both mixtures, but each mixture has a distinct set of two unknown contributors.

As the two mixtures share no unknown contributors under either hypothesis, the peak heights for the mixtures are independent and therefore the joint likelihood ratio is simply the product of the likelihood ratios obtained from the separate analyses. This yields an overwhelmingly strong evidence at 24.6 bans. Note that since K_3 is here replaced by two independent unknown contributors under the defence hypothesis, the upper bound (1.6) does not apply. However, the argument leading to (1.6) is easily extended to establish the higher upper bound of $-2\log_{10}\mathrm{P}(U = K) = 29$ bans, based on the probability $\mathrm{P}(U_1 = K_3, U_2 = K_3)$ that both of the two independent contributors have the profile K_3.

As a side remark, let us informally consider the value of the maximised likelihood under the hypotheses above. The maximised likelihoods $\log_{10}\hat{L}(H_p)$ and $\log_{10}\hat{L}(H'_p)$ under the two different prosecution hypotheses are virtually equal; they both take a value of -248.1. In contrast, the maximised likelihoods under the defence hypothesis are very different: $\log\hat{L}(H_d) = -262.2$ and $\log\hat{L}(H'_d) = -272.6$, which renders H'_d highly unfavourable to the defence. •

5.2.2 Qualitative peak information

The qualitative model presented in Section 2.5 enables an analysis based instead on discrete peak height information. It is unclear whether there is generally enough information to estimate the model parameters under the qualitative model using the likelihood function (2.14); as an alternative we can use estimates based on the full peak height information.

Balding (2013) suggests a model for replicates, which is also based only on the observed discrete peak height information. The model relies on a pre-classification

of alleles and is in that respect also using the peak height information. The weight of evidence, which can be computed through the software likeLTD, is based on maximised likelihood ratios using a penalised likelihood.

Example 5.8 (WoE using the qualitative model). In Cowell et al. (2015), we presented a joint analysis of MC15 and MC18, informally comparing our DNA mixture model with that of Balding (2013).

Considering the similarity of the two mixtures, reflected for instance by their similar EPGs in Figures 1.10-1.12, it is not unreasonable to model them as replicates, which would be the underlying assumption of likeLTD.

To make a similar assumption in our model, we assume that the two mixtures have common mixture proportions ϕ, and further that they have common parameters η and ξ, since these two parameters relate to the laboratory analysis.

We use the set of UK-Caucasian allele frequencies from likeLTD with an F_{ST}-correction of $\theta = 0.02$ and a sampling-adjustment (Balding, 2013) that corresponds to adding the alleles of defendant K_3 to the database. These adjustments effectively increase the probability of the profile K_3, with the effect that the weight of evidence against K_3 is weakened. The maximal WoE – the log of the inverse match probability – against K_3 is reduced correspondingly to 13.2 bans.

We follow Cowell et al. (2015) and use $H_p(3) : K_1 \& K_2 \& K_3$ and $H_d(3) : K_1 \& K_2 \& U_1$ to evaluate the WoE against K_3. When basing the analysis on the observed peak heights, we obtain a WoE of 12.7 bans. Using the likelihood function (2.14) based on only the observed presence and absence of peaks, but in combination with parameter estimates based on the full peak height information, reduces the evidence to 10 bans. This should be compared to a WoE of 8.8 bans, obtained using likeLTD.

Interestingly, for these particular hypotheses we are able to maximise the likelihood function for the qualitative model with the resulting estimates seen in Ta-

ble 5.5. Thus basing the evidence entirely on the qualitative model resulted in a further weakening of the evidence to 9.2 bans.

Table 5.5: Maximum likelihood estimates for $H_p(3)$ and $H_d(3)$ based solely on the qualitative model for peak heights.

Hypothesis	Mixture	μ	σ	ξ	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}	ϕ_{U_1}
$H_p(3)$	MC15	598	0.304	0.138	0.697	0.081	0.222	
	MC18	795	0.304	0.138	0.697	0.081	0.222	
$H_d(3)$	MC15	384	0.276	0.310	0.435	0.152		0.413
	MC18	495	0.276	0.310	0.435	0.152		0.413

The estimates under the two hypotheses are quite different, which was not the case when using the full peak heights, but as the main point of interest is here the value of the maximised likelihood, the quality of estimates is not important.

However, for other choices of hypotheses the likelihood function does not seem to attain its maximum; at least the numerical maximisation exhibits problems with $\rho \to \infty$ and $\eta \to 0$. This highlights the suspicion that there will generally not be enough information in the mixtures to estimate the parameters based solely on peak presence and absence. In contrast, when the estimation is based on the peak heights, a maximum has successfully been found regardless of the specified hypothesis.

A penalised maximum likelihood, which is equivalent to adding prior information about the parameters, may allow the estimation of the parameters based only on absence and presence. This would correspond to the approach of Balding (2013). •

5.3 Prediction of DNA profiles

An important aspect of analysing a DNA mixture lies in determining the composition in terms of the number of contributors, their DNA profiles, and the proportion of DNA from each of the donors. The latter is important in assuring that a contribution is not essentially vanishing: For a DNA sample containing a finite number N of cells,

mixture proportions smaller than $1/N$ can be loosely interpreted as the donor having contributed less than 1 cell.

Below we focus on the prediction of DNA profiles of the contributors under a given hypothesis H. The hypothesis specifies a set of known and/or unknown contributors to the sample and thus the maximal number of contributors is here fixed. Other model parameters ψ are also fixed, but possibly estimated under H. As the profiles of any known contributors are indeed fixed at known configurations, the point of interest is to investigate how the DNA profiles of unknown contributors are explained under the hypothesis H.

We use the posterior distribution $\mathrm{P}_\psi(\boldsymbol{n} \mid H, E)$ given the observed peak height information to rank the possible allocations of DNA profiles. A consequence of using a prior distribution for DNA profiles is that once we establish a set of DNA profiles with a total probability p, there is a probability p that the true profile of the unknown contributor is among these genotypes.

Under the standard model for DNA profiles, where genotypes are prior independent between markers, the genotypes are also independent between markers given the observed evidence. This follows directly from the independence between markers discussed page 70. As a consequence, we can focus on the prediction of the genotypes at a single marker. If desired, these can then be combined into predictions of the full DNA profiles using that the probability of a DNA profile is the product of the probabilities of the individual genotypes.

If some markers carry little information about the genotypes, a list of the most probable configurations of full DNA profiles will reflect this, and we can then instead try to identify a subset of markers where the genotypes are determined with high certainty; a partial profile constructed this way may contain enough information to identify a person.

Cowell et al. (2015) presented an example of predicting a single unknown contributor in the model $H_p(3) : K_1 \,\&\, K_2 \,\&\, U_1$. It was pointed out that the joint analysis of

two mixtures with shared contributors increased the information about the identity of the contributors. In Example 5.9 below, we attack the more complex problem of predicting the profiles of all four contributors, mimicking a situation where no reference profiles are available.

Example 5.9 (Deconvolution of MC15 and MC18). Consider the hypothesis $H : U_1 \,\&\, U_2 \,\&\, U_3 \,\&\, U_4$ for the joint model of MC15 and MC18 and parameters estimated by MLE under H (Table 5.3). Table 5.6 shows the best configuration of DNA profiles measured in terms of highest posterior probability. The configuration has a posterior probability of $1.65 \cdot 10^{-15}$, which is the product of the marker-wise probabilities given. We note that the predicted profile of U_1 coincides with that of K_1.

Table 5.6: The best configuration of DNA profiles under the hypothesis $H : U_1 \& U_2 \& U_3 \& U_4$. A dropped out allele is marked by D.

Marker	U_1	U_2	U_3	U_4	Prob.
D16S539	12,12	11,13	11,12	11,12	0.062
D18S51	14,16	12,14	14,16	15,D	0.009
D19S433	13,14	14,15	14,16.2	13,14	0.037
D21S11	30,31	28,30	28,30	29,30	0.042
D2S1338	23,24	16,17	17,24	D,D	0.035
D3S1358	15,18	17,19	15,17	16,17	0.053
D8S1179	13,16	10,11	13,14	12,13	0.051
FGA	24,26	20,23	21,22	22,D	0.022
TH01	7,8	7,9.3	9,9.3	9.3,9.3	0.042
VWA	14,18	15,19	16,16	16,17	0.022
Joint probability					$1.65 \cdot 10^{-15}$

•

5.3.1 Identifying highest-probability sets

Let H be the hypothesis under consideration, E the included evidence in terms of peak height information (qualitative or quantitative) and known contributors, and let $\boldsymbol{n}$ denote the full set of allele counts – possibly limited to a set of contributors

and alleles of interest. Below we discuss how to identify a set containing a number of configurations $\boldsymbol{n}$ of highest posterior probability $P(\boldsymbol{n} \mid H, E)$.

Although we can obtain the exact probability of any configuration $\boldsymbol{n}$, the state space of interest would generally be too large to fully explore, so tabulating the distribution is not feasible. Nilsson (1998) describes a method for finding the M jointly most probable configurations of *all* variables in a Bayesian network, but we note that this method is not applicable here, as we are interested in most probable configurations of a *subset* of the network variables.

One general way of locating a highest-probability set of states $\boldsymbol{n}$ is by repeated sampling from the relevant distribution $p(\boldsymbol{n})$; we can readily obtain independent samples from any such distribution by Bayesian network techniques. Sampling without replacement would lead to the space being explored faster and would technically be possible (Cowell, 1997), though an implementation of the sampling method is not readily available.

From any set of configurations we can identify a set of most probable configurations as follows. The set $\mathcal{N}$ of distinct configurations accounts for a probability $p(\mathcal{N})$ of the sample space. Any configuration not in $\mathcal{N}$ can therefore at most have a probability $1 - p(\mathcal{N})$. Some of the configurations in $\mathcal{N}$ may also have a probability below $1 - p(\mathcal{N})$, and by removing these we form a smaller set $\mathcal{N}' = \{\boldsymbol{n}' \in \mathcal{N} \ : \ p(\boldsymbol{n}') \geq 1 - p(\mathcal{N})\}$. Now any profile in $\mathcal{N}'$ has probability at least $1 - p(\mathcal{N})$, and any profile outside $\mathcal{N}'$ has a probability smaller than $1 - p(\mathcal{N})$. Consequently, $\mathcal{N}'$ constitutes the $|\mathcal{N}'|$ most probable combinations. Furthermore, $\mathcal{N}'$ contains all configurations with a probability above $1 - p(\mathcal{N})$.

The number of best configurations obtained from $\mathcal{N}$ evidently depends on how many best configurations are contained in $\mathcal{N}$. Since we obtain $\mathcal{N}$ by sampling (with replacement), different criteria for when to stop the sampling will result in different properties of $\mathcal{N}$. Below we discuss three possible criteria.

Number of samples. The simplest criterion sets a fixed number of samples, which will result in a random lower threshold $1 - p(\mathcal{N})$ for probabilities and therefore a random number of best configurations – possibly even none – in the set $\mathcal{N}'$. The runtime is fixed, but there is no guarantee of obtaining a useful list.

Minimum probability. We rely on the method provided by HUGIN through the function `map.configuration` in RHugin. This method terminates the sampling when a minimum probability mass $1 - p_{\min}$ of the state space has been explored, i.e. $1 - p(\mathcal{N}) < p_{\min}$.

The obtained set of configurations is therefore guaranteed to contain all configurations of probability greater than the specified $p_{\min}$. The number of such configurations depends on $p_{\min}$, but is deterministic.

A disadvantage of this approach is that the number of samples N needed is random, so in cases where the distribution of $\boldsymbol{n}$ is diffuse, the runtime can be very long, in particular in combination with a small minimum probability $p_{\min}$. A further slight practical disadvantage is that it requires the user to establish a suitable threshold $p_{\min}$ that ensures the existence of at least one state with probability $p(\boldsymbol{n}) > p_{\min}$.

Minimum number of configurations. If the sampling is stopped when the set of best configurations have reached a pre-determined size N – or when the total probability have reached one – we get a random cut-off both for the probabilities and for the number of samples. In practice, it can be difficult to establish whether N configurations exist because of numerical issues with determining whether a given set has probability exactly 1.

5.3.2 Summarising the DNA profile distribution

The (discrete) space of genotypes, as represented by $\boldsymbol{n}$, is large and thus difficult to explore, even when focusing on a single marker. To extract information, we may focus on the prediction of various subsets of the allele counts.

Prediction of a subset of contributors. It can be a disadvantage to consider the jointly most probable genotypes for all of the unknown contributors: If the posterior is diffuse for some of the contributors, this will lead to the list of most probable configurations of genotypes containing only minor perturbations of genotypes.

Under the standard model for DNA profiles, the posterior distribution is symmetrical in the allocation of genotypes to unknown contributors with equal contributions. Thus, we will not get a unique allocation in the event of perfectly balanced mixtures where also the estimated mixture proportions reflect the balance. In this case, it may be useful to note that the genotypes for single contributors are marginally identically distributed.

Grouping alleles. If there is a significant amount of dropout, it can be an advantage to consider the prediction of groups of alleles. One simple but useful grouping is obtained by collecting all unobserved alleles into one group, which can then be considered a compound allele D. In the further prediction of genotypes, a person can now be assigned allele D on par with any of the observed alleles. Such an assignment denotes that the person possesses one of the alleles in D. This was exploited in Example 5.9.

The idea of considering a group of unseen alleles, is similar to the use of a compound allele in the analyses of Puch-Solis et al. (2013), although in our case we formally marginalise over all the unseen alleles, and therefore do not introduce any approximations.

A related grouping of interest collects all observed alleles into one, in order to further investigate which alleles may have dropped out.

Example 5.10. Under a model with four unknown contributors, separate analyses for MC15 and MC18 point to equal mixture proportions for U_2, U_3, and U_4. The posterior distribution is therefore symmetric in the three contributors, resulting in a flat posterior distribution of genotypes. However, by considering the marginal distribution of U_1 we find that this profile is completely determined and matches the profile of K_1. This conforms with Example 5.7, which gave a maximal WoE against K_1.

For the joint hypothesis for MC15 and MC18 in Example 5.9 involving four unknown contributors, a further investigation of the marginal distributions of each contributor in turn reveals that also here the profile of U_1 is completely determined. The profiles of the other contributors are less well determined, conforming well with the fact that their contributions to the sample are smaller; in particular, dropout is more commonly predicted for the minor contributors. We refrain from reporting these detailed analyses. •

For further examples of summarising the posterior distribution of genotypes, we refer to Section 7.3.2, where the methodology is illustrated using DNAmixtures.

5.4 Identifying stutter and dropout

Given a hypothesis H and a specific allele a, a peak is observed ($Z_a \geq C$) or it is not ($Z_a = 0$) with some probability. Further, the allele a has some probability of being present in the mixture, depending on whether any of the contributors in H possess allele a.

Under our model, we have that an observed peak at a is *due to stutter* if no contributors possess allele a. Similarly allele a has *dropped out* if at least one con-

tributor possesses allele a, but no peak has been observed above threshold. Stutter and dropout may thus be thought of as "false positives" and "false negatives" in the representation of the composition of the mixture by the EPG.

As our model implies a joint model for genotypes, artefacts, and the observed evidence, we can address a multitude of interesting questions regarding the classification of peaks.

Posterior distribution of artefacts. In the analysis of a specific DNA profile, it can be of interest to consider the explanation of a particular peak given the observed peak heights, or possibly just conditionally on partial information about the observed peaks. Of course, if all genotypes are fixed and known, the explanation of any peak is completely determined.

The presence or absence of the allele a can be represented explicitly by a variable Y_a defined as

$$Y_a = \begin{cases} 1, & \sum_i n_{ia} > 0 \\ 0, & \sum_i n_{ia} = 0. \end{cases}$$

For an allele a where a peak has been observed, the probability that the peak is due to stutter alone is now $\mathrm{P}(\text{stutter} \,|\, \boldsymbol{z}, H) = \mathrm{P}(Y_a = 0 \,|\, \boldsymbol{z}, H)$, whereas if no peak has been observed at a, the probability that the allele a has dropped out is $\mathrm{P}(\text{dropout} \,|\, \boldsymbol{z}, H) = \mathrm{P}(Y_a = 1 \,|\, \boldsymbol{z}, H)$.

A set of variables Y_a can readily be included in the network as nodes with parents n_{ia}, after which both stutter- and dropout-probabilities can be obtained by probability propagation, cf. Section 4.3.

Conditional distribution given artefacts. A pre-classification of some alleles as, say, stutter or dropout corresponds to conditioning on specific values of Y_a. Such conditioning can be obtained by a single probability propagation and, should this be desired, a revised analysis can then be performed conditionally on the pre-

classification. Such preprocessing of data would be consistent with our general model specification.

Chapter 6

Challenging the interpretation of a mixture

Any statistical model involves various assumptions and it is important to investigate whether the adopted model adequately describes the specific case at hand. In particular, an effort should be made to justify that each of the hypotheses under consideration represents a plausible explanation of the mixture. A consequence of working with a fully specified statistical model is that we can formally challenge model assumptions. In the following we develop a few examples of diagnostic tools that can reveal problematic aspects of a particular interpretation of a DNA mixture.

In the aim to illustrate the numerous possibilities for reasoning within the scope of the model, we challenge various assumptions made in the analysis of MC15 and MC18. This should not be seen as an attempt to reflect the current practice in case work, but rather as an attempt to highlight a statistical way of reasoning, which may prove useful in practical applications.

6.1 Uncertainty of estimated quantities

The first aspect we address is that of the uncertainty of estimated quantities arising from the variability of the peak heights.

When inference is based on a model where the MLE is used, a sensitivity analysis can be carried out to see how the conclusions are effected by the general uncertainty about the parameters. Examples of such an analysis were presented in Graversen and Lauritzen (2013).

6.1.1 Simulated MLE

Parametric bootstrap – simulation under the model – was used in Graversen and Lauritzen (2013) for assessing the variability in both the MLE and quantities that are functions of the MLE. Examples of such quantities are the likelihood ratios and the probability of a particular profile found by mixture deconvolution.

Given a hypothesis H and model parameters ψ, we repeat for $i = 1, \ldots, N$ the following two steps to obtain samples $\psi_1^*, \ldots, \psi_N^*$ from the distribution $\Pr_\psi(\hat{\psi} \,|\, H)$.

1. Simulate a set of peak heights $\boldsymbol{Z}^i$ under $f_\psi(\boldsymbol{Z} \,|\, H)$.
2. Basing the likelihood on $\boldsymbol{Z}^i$, maximise to find the MLE ψ_i^* – possibly under a hypothesis different to H.

For any function $h(\hat{\psi})$ of the MLE, a sample from the distribution $\Pr_\psi(h(\hat{\psi}) \,|\, H)$ can now be obtained as $h(\psi_1^*), \ldots, h(\psi_N^*)$.

If the aim of the analysis is to assess the variability in the MLE on a particular case, where the true ψ is not known, we simulate under the model replacing ψ by the estimate $\hat{\psi}$ based on the observed set of peak heights.

A drawback of the simulation approach is that the likelihood needs to be maximised repeatedly, which is computationally expensive for models with many unknown contributors.

Example 6.1 (Simulated MLEs under $H_p(4)$, MC15). Using the MLE under $H_p(4)$ for MC15, we sample 2000 new sets of peak heights and estimate the parameters on each. As a starting point for the maximisations we also use the observed MLE. Not all maximisations were successful in the first attempt; in 44 of the 2000

cases the encounter of an implausible set of parameters caused the propagation of evidence during the evaluation of the likelihood function to fail. For these cases, we have restarted the maximisation using two alternative starting points. Inspection did not suggest a systematic connection between the failed maximisation and the location of the parameters and is more likely due to the specific optimisation method.

The computation time for the 2000 maximisations was around 5 hours – one maximisation taking on average 9 seconds – performed on a standard desktop using the current version of DNAmixtures.

Figure 6.1 reflects the approximate inversely proportional relationship (5.1) between $\hat{\rho}$ and $\hat{\eta}$ described in Section 5.1. In contrast, the parameters μ and σ are fairly orthogonal, which is one advantage of using this parameterisation. In particular, the marginal confidence intervals based on this parameterisation better reflect the joint range of plausible parameter values.

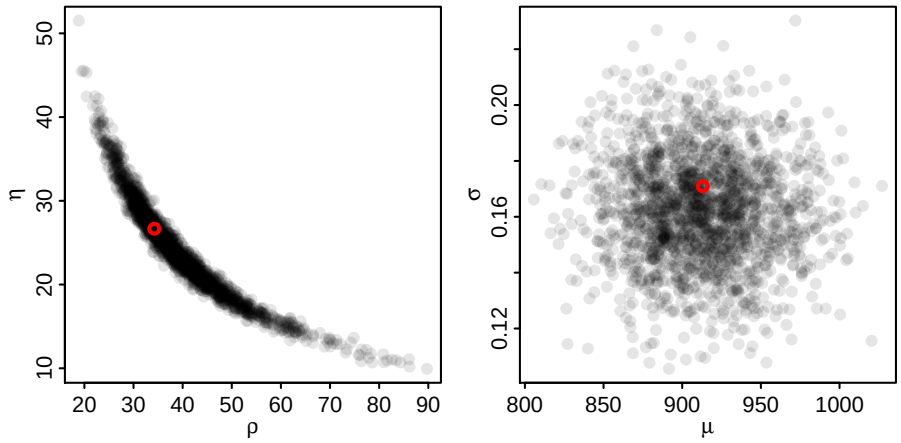


Figure 6.1: Simulated values of $(\hat{\rho}, \hat{\eta})$ and $(\hat{\mu}, \hat{\sigma})$ under the model $H_p(4)$ for MC15 using the observed MLE (red) as the true parameter.

The distributions of $\hat{\rho}$ and $\hat{\eta}$ are noticeably skewed, as seen more clearly from Figure 6.2. In contrast, the normal approximation seems fine for both $\hat{\mu}$ and $\hat{\sigma}$.

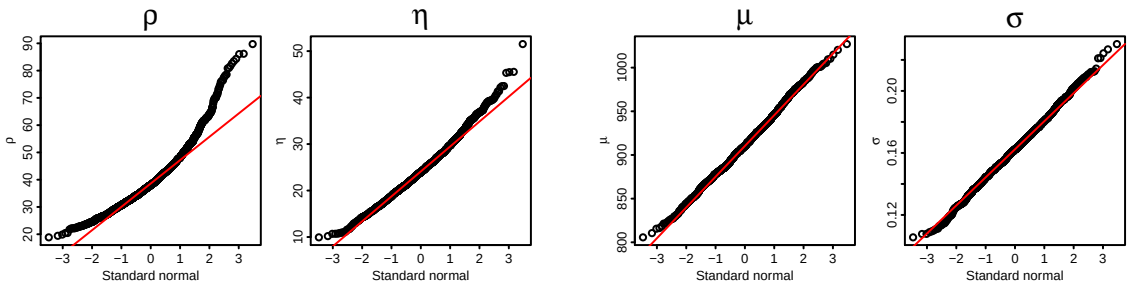


Figure 6.2: Comparing the quantiles of simulated MLEs to those of the standard normal distribution.

The mixture proportions are well determined, as is apparent from Figure 6.3 showing the concentrated distribution of their estimates over the simplex.

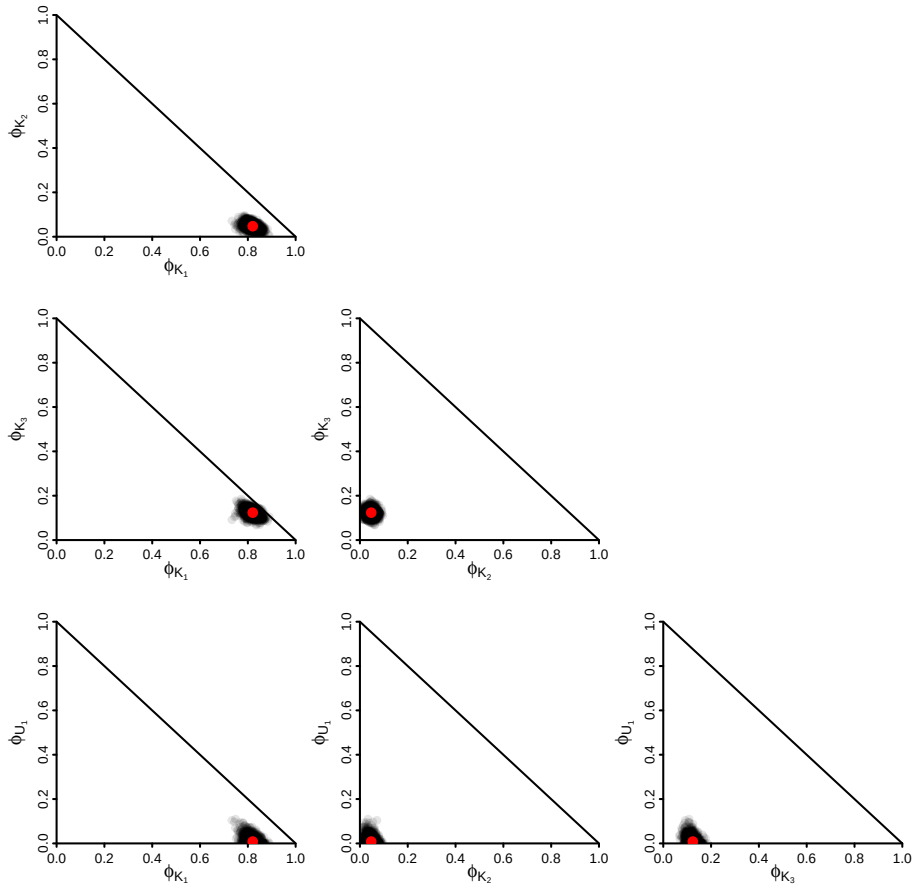


Figure 6.3: Simulated MLE for mixture proportions ϕ under $H_p(4)$. The observed MLEs $\hat{\phi}_{U_1} = 0.008$, $\hat{\phi}_{K_1} = 0.821$, $\hat{\phi}_{K_2} = 0.047$, and $\hat{\phi}_{K_3} = 0.124$ are marked in red.

A closer look at the estimates reveals that the estimated mixture proportion for the unknown contributor essentially has a point mass in zero (Figure 6.4), rendering

the normal approximation unsuitable for $\hat{\phi}_{U_1}$. Some estimates for ϕ_{K_2} are also zero. The estimates for other mixture proportions are well within the parameter space and their marginal distributions are close to normal, as is the case for the estimated stutter proportion $\hat{\xi}$; we omit the supporting plots.

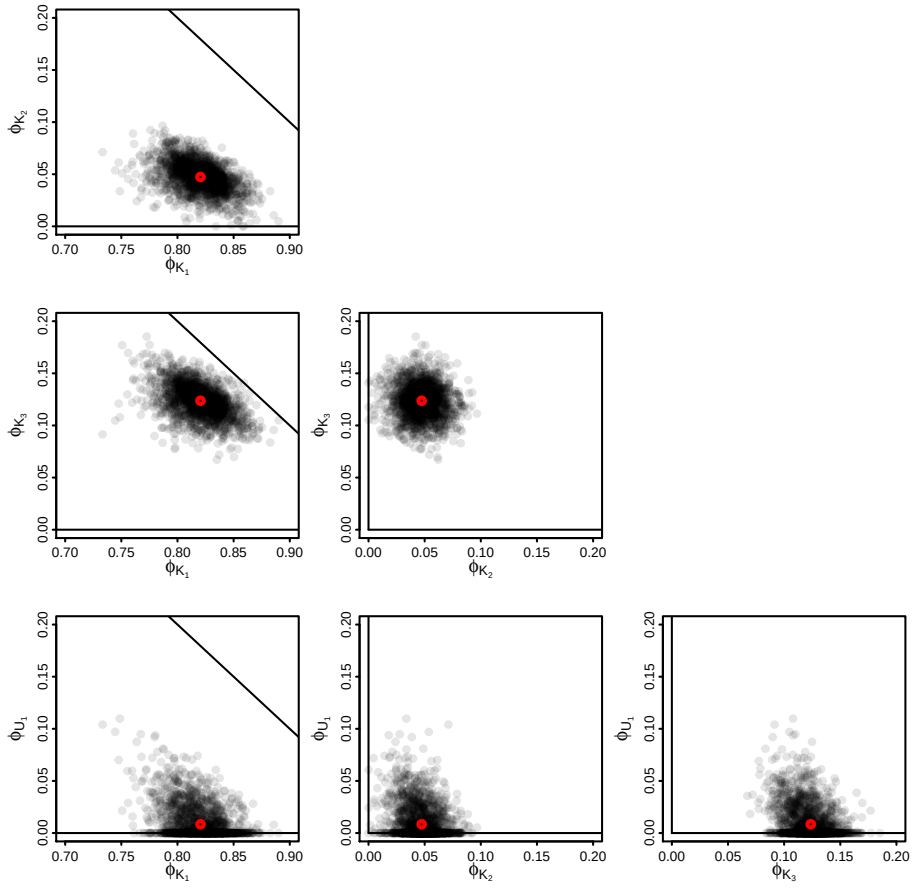


Figure 6.4: Simulated MLE for mixture proportions ϕ. Magnified version of Figure 6.3.

•

Example 6.2 (Simulated MLE under $H_d(4)$, MC15). For the model $H_d(4)$ there are two unknown contributors, and the MLE lies on the boundary with $\hat{\psi}_{U_1} = \hat{\psi}_{U_2} = 0.081$. We have followed the procedure in Example 6.1 and sampled 2000 MLEs using the observed MLE as the true parameter, with 34 of the maximisations restarted from one or two different starting points. Performing the 2000 maximisations in this model with two unknown contributors took around 8.5

hours – on average 15 seconds per maximisation – compared to a total of 5 hours under the model in Example 6.1 involving a single unknown contributor.

The estimated mixture proportions seen in Figure 6.5 are again well determined. The shaded areas show how the order restrictions on the proportions for unknown contributors affect the parameter space.

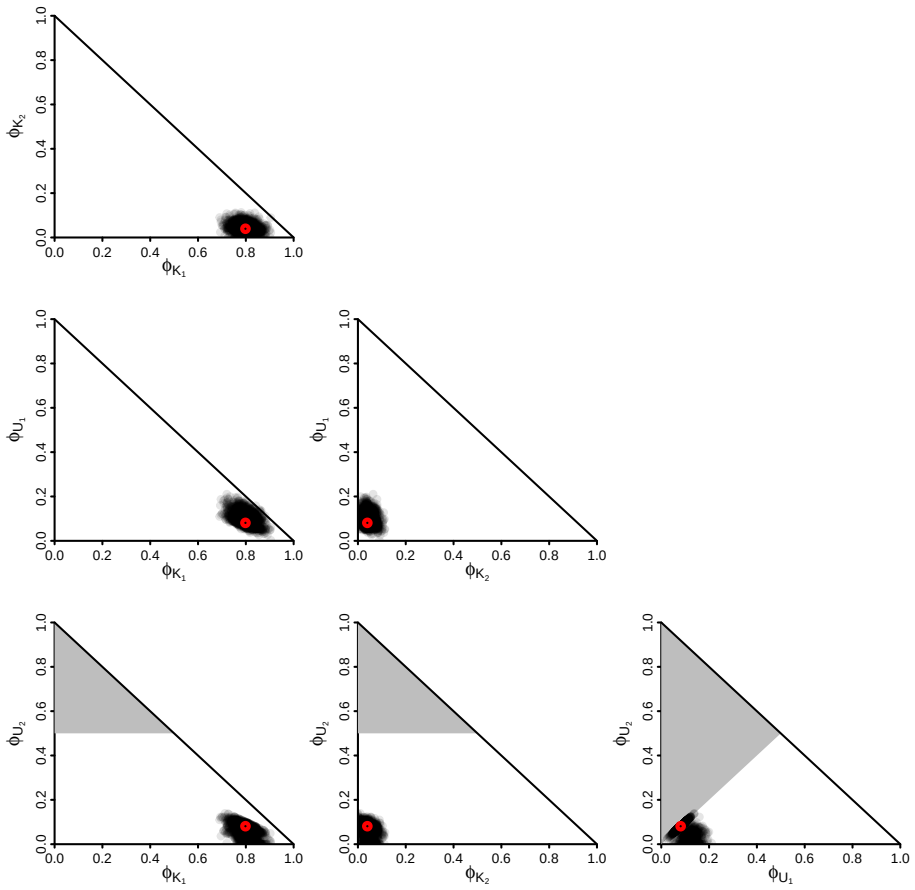


Figure 6.5: MLE for mixture proportions ϕ simulated under $H_d(4)$ for MC15. The observed MLE is marked in red. The shaded areas mark the restrictions of the parameter space due to the ordering $\phi_{U_1} \geq \phi_{U_2}$.

The more detailed plot in Figure 6.5 reveals that many of the estimates lie on the boundaries $\phi_{U_1} = 0$, $\phi_{U_2} = 0$, $\phi_{K_2} = 0$, or $\phi_{U_1} = \phi_{U_2}$. A few of the estimates for ξ are also zero (not shown).

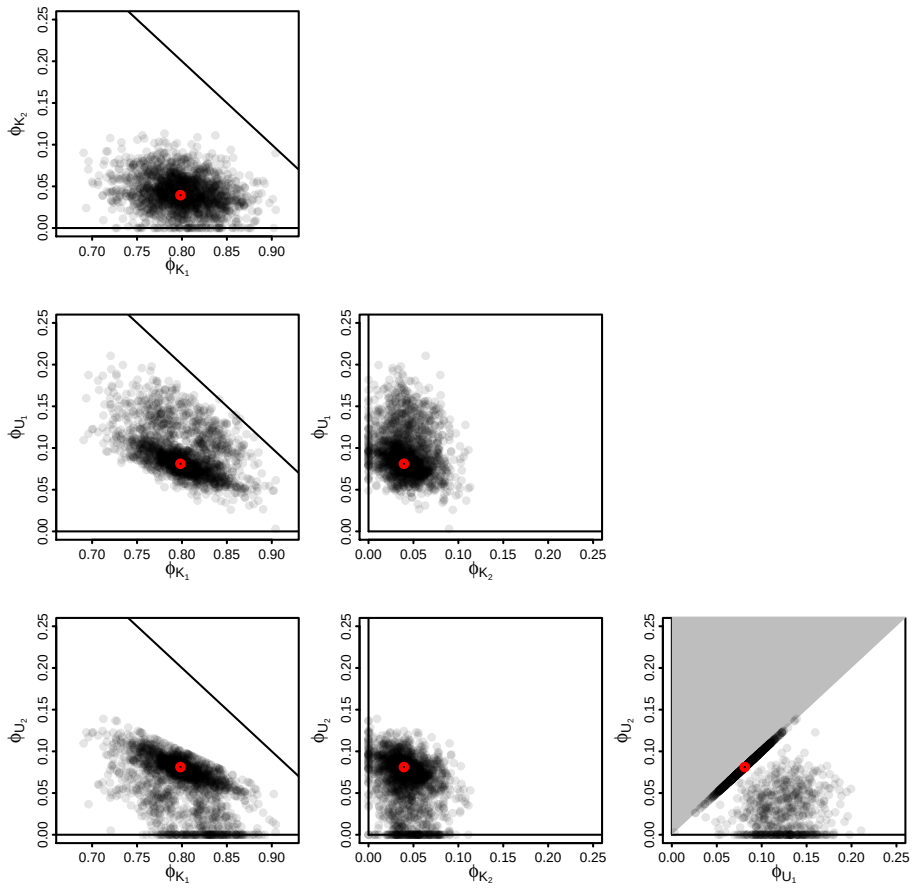


Figure 6.6: MLE for mixture proportions ϕ simulated under $H_d(4)$ for MC15. The observed MLE is marked in red. The shaded areas mark the restrictions of the parameter space due to the ordering $\phi_{U_1} \geq \phi_{U_2}$. Magnified version of Figure 6.5.

•

6.1.2 Asymptotic variance for the MLE

Another possibility for assessing the variability of the estimated quantities is to rely on the asymptotic normality of the MLE, although care should be taken when parameters are on or close to the boundary. We use the inverse of the observed information matrix evaluated in the MLE as an estimate of the asymptotic variance of the MLE.

Firstly, we derive the variance matrix in the case of modelling a single DNA mixture. As the mixture proportions sum to 1, any of the parameters is completely

determined by the $k-1$ other parameters. In order for the observed information to be of full rank we parametrise using only $k-1$ of the mixture proportions $\phi_1, \ldots, \phi_{k-1}$, thus omitting $\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$.

We assume that the true parameter (and the observed MLE) lie in the interior of the parameter space described in Section 5.1, i.e. that $0 < \rho$, $0 < \eta$, $0 < \xi < 1$, and finally $0 < \phi_i < 1, i = 1, \ldots, k-1$ with the constraint that $\sum_{i=1}^{k-1} \phi_i < 1$.

When order constraints are imposed on the mixture proportions, the parameter lies in the part of the space also satisfying $\phi_i \leq \phi_j$ for unknown contributors $i > j$. Note that for all interior points ($\phi_i < \phi_j$) in the restricted parameter space, the observed information matrix, and thus the variance matrix, is the same as when the order constraints are not imposed. We return in Example 6.4 to the special case of equal mixture proportions for unknown contributors.

We find derivatives by numerical methods, and so be able to perform the differentiation in an unconstrained parameter space, we use a re-parametrisation

$$(\alpha_1, \alpha_2, \alpha_3, \beta_1, \ldots, \beta_{k-1}) = \left(\log \rho, \log \eta, \log \frac{\xi}{1-\xi}, \log \frac{\phi_1}{\phi_k}, \ldots, \log \frac{\phi_{k-1}}{\phi_k} \right)$$

where now the parameter space is $\mathbb{R}^{3+k-1}$.

The observed observation matrix $\tilde{H} = -\,\mathrm{D}^2 \log \tilde{L}(\alpha, \beta)$ is evaluated in the found MLE $(\hat{\alpha}, \hat{\beta})$ by numerical differentiation using `hessian` from the numDeriv R package. As a side remark, note that this computed Hessian of the log-likelihood function can also be used in checking that the postulated MLE is indeed a maximum point.

By inverting the observed information matrix, we obtain the variance matrix $\tilde{H}(\hat{\alpha}, \hat{\beta})^{-1}$ for $(\hat{\alpha}, \hat{\beta})$. The variance matrix for $(\hat{\rho}, \hat{\eta}, \hat{\xi}, \hat{\phi})$ is found by the delta method as

$$\hat{\mathbb{V}}(\hat{\rho}, \hat{\eta}, \hat{\xi}, \hat{\phi}) = \mathrm{D}\, h(\hat{\alpha}, \hat{\beta})^T \tilde{H}^{-1}(\hat{\alpha}, \hat{\beta})\, \mathrm{D}\, h(\hat{\alpha}, \hat{\beta}),$$

using that

$$\begin{aligned}(\rho, \eta, \xi, \phi_1, \ldots, \phi_{k-1}) &= h(\alpha_1, \alpha_2, \alpha_3, \beta_1, \ldots, \beta_{k-1}) \\ &= \left(e^{\alpha_1}, e^{\alpha_2}, \frac{e^{\beta_3}}{1+e^{\beta_3}}, \frac{e^{\beta_1}}{1+\sum_{i=1}^{k-1} e^{\beta_i}}, \ldots, \frac{e^{\beta_{k-1}}}{1+\sum_{i=1}^{k-1} e^{\beta_i}}\right).\end{aligned}$$

The Jacobian of h is a block diagonal matrix with one block for each of the parameters ρ, η, ξ, and ϕ. The first three blocks are

$$\frac{\partial \rho}{\partial \alpha_1} = e^{\alpha_1} = \rho, \qquad \frac{\partial \eta}{\partial \alpha_2} = e^{\alpha_2} = \eta, \qquad \frac{\partial \xi}{\partial \alpha_3} = e^{\alpha_3} - (e^{\alpha_3})^2 = \xi - \xi^2,$$

For the $(k-1) \times (k-1)$-block containing derivatives of the mixture proportions, the diagonal elements are

$$\frac{\partial \phi_i}{\partial \beta_i} = \frac{e^{\beta_i}}{1+\sum_l e^{\beta_l}} - \left(\frac{e^{\beta_i}}{1+\sum_l e^{\beta_l}}\right)^2 = \phi_i - \phi_i^2,$$

and the off-diagonal elements ($i \neq j$) are

$$\frac{\partial \phi_i}{\partial \beta_j} = -\frac{e^{\beta_i}}{1+\sum_l e^{\beta_l}} \frac{e^{\beta_j}}{1+\sum_l e^{\beta_l}} = -\phi_i \phi_j.$$

The k'th mixture proportion $\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$ is easily included in the covariance matrix, but note that the resulting covariance matrix will be singular. Similarly, the asymptotic variance for the MLE in the parametrisation using $\mu = \rho\eta$ and $\sigma = 1/\sqrt{\rho}$ can be obtained by the delta method.

When some parameters are fixed at known values, rows and columns corresponding to the fixed parameters are simply omitted from the observed information matrix.

When multiple mixtures are analysed jointly, for simplicity we allow each of the four parameters to be either 1) fixed and known for all samples in the model, 2) common across all samples, or 3) completely unrestricted. Extending from the case of a single mixture is then straightforward: In the first and second case, the parameters collapse into one, after when the partial derivative is as above for the single mixture, and in the third case the partial derivatives are simply 0 between

mixtures. Note that if the mixtures analysed have no unknown contributors in common, then their respective sets of parameter estimates are uncorrelated.

As there is a one-to-one correspondence between ρ and σ, σ is fixed or common across samples exactly when ρ is. Equality constraints on ρ and η translate to constraints on $\mu = \rho\eta$, but how they translate depends on the specific constraints.

Example 6.3 (Asymptotic variance for the MLE, MC15). Table 6.1 gives the estimated correlation matrix and standard errors for the maximum likelihood estimates under $H_p(4)$ for MC15. Unsurprisingly, the mixture proportions are negatively correlated as they add up to 1. Conforming with Figure 6.1, we see that $\hat{\rho}$ and $\hat{\eta}$ have a correlation of -0.983, whereas $\hat{\mu}$ and $\hat{\sigma}$ have a correlation of only -0.029. The estimated coefficient of variation for $\hat{\mu}$ is 3.8% and indicates that μ is well determined by the mixture, whereas σ is less so with a coefficient of variation of 10% for $\hat{\sigma}$.

Table 6.1: Correlation matrix and standard errors based on asymptotic normality of the MLE. Mixture proportions (grey) are negatively correlated. There is a strong correlation between $\hat{\rho}$ and $\hat{\eta}$ (red), but not between $\hat{\mu}$ and $\hat{\sigma}$ (blue).

	ρ	η	μ	σ	ξ	ϕ_{U_1}	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}
ρ	1.000	-0.983	0.029	-1.000	-0.064	0.088	0.050	-0.030	-0.144
η	-0.983	1.000	0.153	0.983	0.065	-0.078	-0.055	0.029	0.140
μ	0.029	0.153	1.000	-0.029	0.009	0.051	-0.033	-0.003	-0.016
σ	-1.000	0.983	-0.029	1.000	0.064	-0.088	-0.050	0.030	0.144
ξ	-0.064	0.065	0.009	0.064	1.000	-0.311	0.212	0.031	0.070
ϕ_{U_1}	0.088	-0.078	0.051	-0.088	-0.311	1.000	-0.478	-0.319	-0.297
ϕ_{K_1}	0.050	-0.055	-0.033	-0.050	0.212	-0.478	1.000	-0.331	-0.442
ϕ_{K_2}	-0.030	0.029	-0.003	0.030	0.031	-0.319	-0.331	1.000	-0.068
ϕ_{K_3}	-0.144	0.140	-0.016	0.144	0.070	-0.297	-0.442	-0.068	1.000
Estimate	34.2	26.7	913	0.171	0.074	0.008	0.821	0.047	0.124
SE	7.131	5.619	35.0	0.018	0.014	0.019	0.020	0.014	0.015

The empirical correlations and standard errors (Table 6.2) are in accordance with those based on the observed information (Table 6.1).

Table 6.2: Empirical correlations and standard errors based on simulations under $H_p(4)$. We see also here a weak correlation (blue) between $\hat{\mu}$ and $\hat{\sigma}$, as well as a negative correlation (grey) between mixture proportions.

	μ	σ	ξ	ϕ_{U_1}	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}
μ	1.000	-0.057	0.060	-0.000	0.013	-0.015	-0.002
σ	-0.057	1.000	0.071	-0.317	0.058	0.074	0.181
ξ	0.060	0.071	1.000	-0.278	0.192	-0.061	0.094
ϕ_{U_1}	-0.000	-0.317	-0.278	1.000	-0.374	-0.237	-0.327
ϕ_{K_1}	0.013	0.058	0.192	-0.374	1.000	-0.460	-0.469
ϕ_{K_2}	-0.015	0.074	-0.061	-0.237	-0.460	1.000	-0.083
ϕ_{K_3}	-0.002	0.181	0.094	-0.327	-0.469	-0.083	1.000
Estimate	913	0.171	0.074	0.008	0.821	0.047	0.124
SE	34.8	0.019	0.016	0.016	0.020	0.014	0.016

•

Example 6.4 (Estimated covariance matrix under equal mixture proportions). Computing a meaningful covariance matrix under the defence $H_d(4)$: K_1 & K_2 & U_1 & U_2 based on MC15 is more complicated than for $H_p(4)$, since the maximum lies on the boundary $\phi_{U_1} = \phi_{U_2}$ (see Example 5.5).

We may lift the order restriction on the unknown contributors and compute the unconstrained variance matrix V for the MLE $(\hat{\mu}, \hat{\sigma}, \hat{\xi}, \hat{\phi}_{U_1}, \hat{\phi}_{U_2}, \hat{\phi}_{K_1})$ in this larger model. Thereafter, we condition on the observed event $\hat{\phi}_{U_1} = \hat{\phi}_{U_2}$ as follows.

Firstly, we compute the variance matrix $\tilde{V}$ for $(\hat{\mu}, \hat{\sigma}, \hat{\xi}, \hat{\phi}_{U_1}, \hat{\phi}_{U_2} - \hat{\phi}_{U_1}, \hat{\phi}_{K_1})$. In the corresponding concentration matrix $\tilde{K} = \tilde{V}^{-1}$, we can condition on $\hat{\phi}_{U_2} - \hat{\phi}_{U_1} = 0$, simply by removing the corresponding row and column; see e.g. (Lauritzen, 1996, pp 256-57). Inverting the thereby obtained concentration matrix gives the conditional covariance matrix given $\hat{\phi}_{U_2} - \hat{\phi}_{U_1} = 0$. Finally, we use the delta method to include $\hat{\phi}_{K_2} = 1 - (\hat{\phi}_{U_1} + \hat{\phi}_{U_2} + \hat{\phi}_{K_1})$ in the covariance matrix. The resulting correlation matrix and standard errors are given in Table 6.3.

Table 6.3: Conditional correlation matrix and standard errors for the MLE under $H_d(4)$ for MC15 given that $\hat{\phi}_{U_1} = \hat{\phi}_{U_2}$.

	μ	σ	ξ	ϕ_{U_1}	ϕ_{K_1}	ϕ_{K_2}
μ	1.000	-0.023	0.027	-0.010	-0.004	0.020
σ	-0.023	1.000	0.098	-0.004	-0.027	0.044
ξ	0.027	0.098	1.000	-0.101	0.185	-0.126
ϕ_{U_1}	-0.010	-0.004	-0.101	1.000	-0.742	-0.301
ϕ_{K_1}	-0.004	-0.027	0.185	-0.742	1.000	-0.416
ϕ_{K_2}	0.020	0.044	-0.126	-0.301	-0.416	1.000
Estimate	914	0.198	0.072	0.081	0.798	0.039
SE	40.6	0.023	0.019	0.013	0.028	0.019

•

Example 6.5 (Asymptotic variance for the joint MLE for MC15 and MC18). For completeness, we also compute the asymptotic covariance matrices for the MLE under each of hypotheses $H_p(4)$ and $H_d(4)$ for the joint analysis of MC15 and MC18. Both correlations and standard errors, seen in Table 6.4, are very similar to those obtained under the separate models above. We notice that although the two mixtures do share unknown contributors – so their peak heights are not independent – the correlations between the two sets of parameter estimates corresponding to the two mixtures are weak. The correlations within a mixture are stronger. Of course, if restrictions are introduced on the parameters – for instance using common η and ξ for the two mixtures as in Cowell et al. (2015) – we can get strong correlations between mixtures.

Table 6.4: Standard errors and correlation matrix based on the asymptotic variance matrix for the joint MLE for MC15 and MC18.

(a) $H_p(4) : K_1$ & K_2 & K_3 & U_1

		MC15							MC18						
		μ	σ	ξ	ϕ_{U_1}	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}	μ	σ	ξ	ϕ_{U_1}	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}
MC15	μ	1.000	-0.029	0.009	0.050	-0.032	-0.002	-0.016	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	σ	-0.029	1.000	0.064	-0.088	-0.050	0.030	0.145	0.000	0.000	0.001	-0.002	0.001	0.000	0.000
	ξ	0.009	0.064	1.000	-0.313	0.213	0.029	0.072	0.000	0.000	0.008	-0.005	0.005	0.000	-0.001
	ϕ_{U_1}	0.050	-0.088	-0.313	1.000	-0.479	-0.312	-0.300	-0.001	0.000	-0.006	0.011	-0.008	-0.003	0.001
	ϕ_{K_1}	-0.032	-0.050	0.213	-0.479	1.000	-0.335	-0.440	0.000	0.000	0.005	-0.006	0.007	-0.001	-0.002
	ϕ_{K_2}	-0.002	0.030	0.029	-0.312	-0.335	1.000	-0.069	0.000	-0.001	-0.002	0.002	-0.002	0.005	-0.004
	ϕ_{K_3}	-0.016	0.145	0.072	-0.300	-0.440	-0.069	1.000	0.000	0.001	0.002	-0.008	0.003	0.000	0.004
MC18	μ	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	1.000	-0.022	0.007	0.042	-0.020	-0.011	-0.009
	σ	0.000	0.000	0.000	0.000	0.000	-0.001	0.001	-0.022	1.000	0.088	-0.098	-0.038	0.073	0.081
	ξ	0.000	0.001	0.008	-0.006	0.005	-0.002	0.002	0.007	0.088	1.000	-0.362	0.154	0.080	0.111
	ϕ_{U_1}	0.000	-0.002	-0.005	0.011	-0.006	0.002	-0.008	0.042	-0.098	-0.362	1.000	-0.394	-0.272	-0.302
	ϕ_{K_1}	0.000	0.001	0.005	-0.008	0.007	-0.002	0.003	-0.020	-0.038	0.154	-0.394	1.000	-0.386	-0.472
	ϕ_{K_2}	0.000	0.000	0.000	-0.003	-0.001	0.005	0.000	-0.011	0.073	0.080	-0.272	-0.386	1.000	-0.134
	ϕ_{K_3}	0.000	0.000	-0.001	0.001	-0.002	-0.004	0.004	-0.009	0.081	0.111	-0.302	-0.472	-0.134	1.000

	μ	σ	ξ	ϕ_{U_1}	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}	μ	σ	ξ	ϕ_{U_1}	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}
Estimate	913	0.171	0.074	0.008	0.820	0.047	0.124	1056	0.166	0.085	0.010	0.706	0.091	0.194
SE	35.0	0.018	0.014	0.018	0.020	0.014	0.015	39.3	0.017	0.016	0.019	0.022	0.016	0.018

(b) $H_d(4) : K_1$ & K_2 & U_1 & U_2

		MC15							MC18						
		μ	σ	ξ	ϕ_{U_1}	ϕ_{U_2}	ϕ_{K_1}	ϕ_{K_2}	μ	σ	ξ	ϕ_{U_1}	ϕ_{U_2}	ϕ_{K_1}	ϕ_{K_2}
MC15	μ	1.000	-0.029	0.009	-0.015	0.048	-0.031	-0.002	0.000	0.000	0.000	0.000	0.002	-0.002	0.001
	σ	-0.029	1.000	0.061	0.126	-0.077	-0.060	0.053	0.000	0.013	-0.005	-0.003	0.006	-0.016	0.019
	ξ	0.009	0.061	1.000	0.075	-0.314	0.210	0.028	0.000	0.018	0.055	0.000	-0.028	0.013	0.013
	ϕ_{U_1}	-0.015	0.126	0.075	1.000	-0.299	-0.434	-0.064	0.000	0.006	0.014	0.031	-0.024	-0.006	0.000
	ϕ_{U_2}	0.048	-0.077	-0.314	-0.299	1.000	-0.481	-0.301	0.001	0.001	-0.032	-0.007	0.076	-0.049	-0.010
	ϕ_{K_1}	-0.031	-0.060	0.210	-0.434	-0.481	1.000	-0.353	-0.001	-0.020	0.017	-0.013	-0.045	0.080	-0.044
	ϕ_{K_2}	-0.002	0.053	0.028	-0.064	-0.301	-0.353	1.000	0.000	0.022	0.003	-0.005	-0.008	-0.046	0.078
MC18	μ	0.000	0.000	0.000	0.000	0.001	-0.001	0.000	1.000	-0.022	0.007	-0.008	0.040	-0.018	-0.010
	σ	0.000	0.013	0.018	0.006	0.001	-0.020	0.022	-0.022	1.000	0.110	0.063	-0.086	-0.061	0.111
	ξ	0.000	-0.005	0.055	0.014	-0.032	0.017	0.003	0.007	0.110	1.000	0.104	-0.358	0.148	0.086
	ϕ_{U_1}	0.000	-0.003	0.000	0.031	-0.007	-0.013	-0.005	-0.008	0.063	0.104	1.000	-0.293	-0.467	-0.130
	ϕ_{U_2}	0.002	0.006	-0.028	-0.024	0.076	-0.045	-0.008	0.040	-0.086	-0.358	-0.293	1.000	-0.398	-0.258
	ϕ_{K_1}	-0.002	-0.016	0.013	-0.006	-0.049	0.080	-0.046	-0.018	-0.061	0.148	-0.467	-0.398	1.000	-0.411
	ϕ_{K_2}	0.001	0.019	0.013	0.000	-0.010	-0.044	0.078	-0.010	0.111	0.086	-0.130	-0.258	-0.411	1.000

	μ	σ	ξ	ϕ_{U_1}	ϕ_{U_2}	ϕ_{K_1}	ϕ_{K_2}	μ	σ	ξ	ϕ_{U_1}	ϕ_{U_2}	ϕ_{K_1}	ϕ_{K_2}
Estimate	913	0.174	0.074	0.123	0.010	0.819	0.049	1056	0.168	0.085	0.194	0.011	0.703	0.092
SE	35.6	0.018	0.015	0.016	0.019	0.021	0.014	39.8	0.017	0.016	0.019	0.019	0.023	0.017

•

Confidence intervals. Approximate confidence intervals can be based on the asymptotic normality of the MLE or a suitable transformation that is closer to normal.

Using instead simulated MLEs, simple confidence intervals

$$[k_{\alpha/2}, k_{1-\alpha/2}] \quad \text{or} \quad [2h(\hat{\psi}) - k_{1-\alpha/2}, 2h(\hat{\psi}) - k_{\alpha/2}],$$

may be constructed directly from the observed $\alpha/2$ and $1-\alpha/2$ percentiles for the quantity $h(\psi)$ of interest. The latter interval is generally better (Davison and Hinkley, 1997) and builds on the assumption that the distributions of $h(\hat{\psi}^*) - h(\hat{\psi})$ and $h(\hat{\psi}) - h(\psi)$ are similar. However, the former interval is guaranteed to respect the restrictions on the parameter-space.

Alternatively, an interval can be based on the profile likelihood for $h(\psi)$, which we may easily evaluate since we can maximise the likelihood for fixed values of $h(\psi)$ by imposing this as a constraint. Such an interval is also guaranteed to be contained within the parameter space, and affords a particularly useful approach in the situation of estimates on the boundary where relying on approximate normality is inappropriate.

6.2 Hypothesis testing

A consequence of working with a fully specified statistical model is that we can challenge the assumptions of the model using established statistical methodology. In the following we illustrate a range of questions that can be addressed by means of likelihood ratio tests.

6.2.1 Assumptions on parameters

Some of the model parameters, such as ξ and η, relate to the laboratory procedure rather than the specific DNA samples; thus, they would be expected to be common for all mixtures in a joint analysis, and making such an assumption would limit the number of unknown parameters. We can formally investigate such propositions by means of likelihood ratio tests.

Example 6.6 (Common parameters for MC15 and MC18). Firstly, we address the assumption of common parameters η and ξ for MC15 and MC18 made in Cowell et al. (2015).

Common η and common ξ. A test for the reduction from $2 \times 2 = 4$ parameters to $2 \times 1 = 2$ parameters can be done in terms of a χ^2-test on 2 degrees of freedom. Under $H_p(4)$, the likelihood ratio statistic is 0.369, which leads to a p-value of 0.98. Thus, the use of common parameters η and ξ for the two mixtures is justifiable. Similarly, the reduction is supported under $H_d(4)$ with a p-value of 0.99.

Common ϕ. In the comparison to likeLTD in Example 5.8, it was assumed that the three individuals had contributed the same proportion of DNA to each of MC15 and MC18. This assumption can also be addressed by a likelihood ratio test.

Generally, setting the mixture proportions to be equal in R mixtures with k contributors reduces the dimension of the parameter space by $(R-1)(k-1)$. In the setting of Example 5.8, there are $R = 2$ mixtures and $k = 3$ contributors and thus 2 degrees of freedom.

Although the estimated mixture proportions may seem similar for MC15 and MC18, a test for their equality is firmly rejected with a p-value of 0.0003 under both $H_p(3)$ and $H_d(3)$. •

6.2.2 The number of contributors

We have so far taken the number of contributors to be known, but in practice determining the number of contributors is part of the analysis of a mixture. There will often be some minimum number of contributors needed to explain a mixture; ignoring the possibility of stutter, if a marker exhibits N peaks there need to be at least $\lceil N/2 \rceil$ contributors. In contrast there is no upper bound, as contributors may share alleles, or alleles can have dropped out.

Note that adding an unknown contributor U to any hypothesis gives a larger model – the smaller model is the special case of $\phi_U = 0$ – so the likelihood is non-decreasing as we increase the number of unknown contributors. Particularly, this means that establishing a suitable number of contributors cannot be done by maximum likelihood.

There are many other possible ways of determining a suitable number of contributors. One approach, which we discuss below, is to find an adequate explanation of the mixture and then consider whether the addition of an unknown contributor is significant. Another possibility would be to introduce a penalty for including many contributors and then maximise over the penalised likelihood. A third solution introduces a prior distribution on the number of contributors. In this case, integrated rather than maximised likelihoods are used for the weight of evidence. For a further discussion, see Cowell et al. (2015).

Example 6.7 (Profile likelihood for the number of contributors). Figure 6.7 shows how the maximised likelihoods increase as more contributors are added to the joint model for MC15 and MC18. We note that in this case the WoE decreases slightly. However, it may also happen that the WoE increases; this is indeed the case for the WoE against K_2 for the mixture MC15 where the WoE increases slowly from -0.08 bans to 0.09 bans when increasing the number of contributors from three to six.

•

Test for zero contributions

In various settings it may be of interest to determine whether one or more mixture proportions are essentially zero. Such tests are possible, though can be difficult due to the many one-sided constraints on mixture proportions. Note in particular that $\phi_{U_i} = 0$ implies that $\phi_{U_j} = 0$ for any $j > i$. Since the null hypothesis lies on the

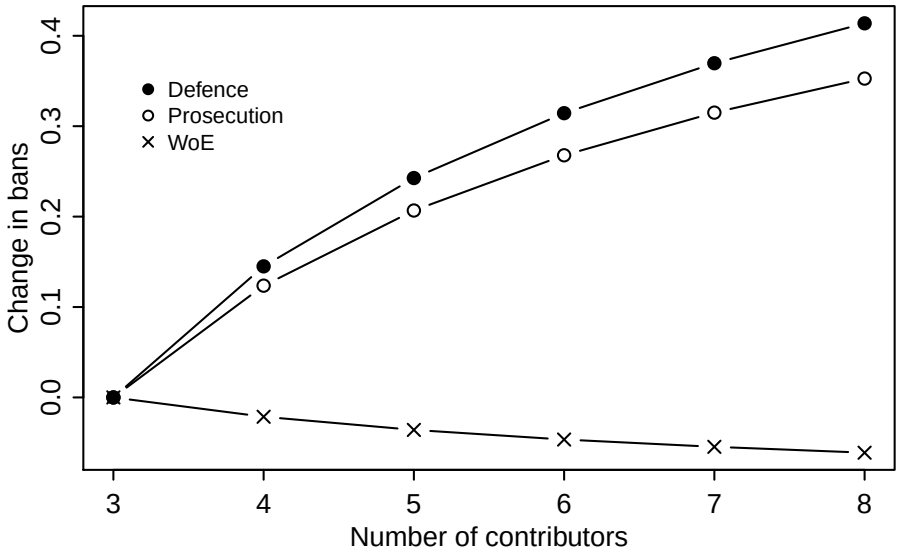


Figure 6.7: Profile likelihood for the number of contributors. The illustration is taken from Cowell et al. (2015).

boundary of the parameter space, the likelihood ratio statistic is not asymptotically chi-square distributed, loosely speaking because the maximum likelihood estimator can only fall on "one side" of the boundary and thus cannot be approximately normal; see e.g. the discussion in (Davison, 2003, section 4.6).

In the setting of El Barmi and Dykstra (1999), we may approach the problem as follows. Assume that we are in a situation where the parameter space is constrained by $h_j(\psi_0) \geq 0, j = 1, \ldots, s$ and we wish to test whether $h_j(\psi_0) = 0, j = 1, \ldots, s$.

Provided that various regularity conditions are satisfied, (El Barmi and Dykstra, 1999, Theorem 3.2) states that the limiting distribution of the likelihood ratio statistic Q is a mixture of chi-square distributions. In particular, a p-value can be computed as

$$\mathrm{P}(Q \geq t) \approx \sum_{j=0}^{s} w_j(\psi_0)\, \mathrm{P}(\chi^2_{s-j} \geq t). \tag{6.1}$$

Here, w_j are weights that are possible though slightly complicated to evaluate.[1] In the simple case of a single constraint h_1, then $w_0(\psi_0) = 1/2$.[2]

[1] Each weight is based on probabilities of the form $\mathrm{P}(\mathcal{N}(\mathbf{0}, \Sigma(\psi_0)) \geq \mathbf{0})$, where $\Sigma(\psi_0)$ is a sub-matrix of the inverse Fisher information in ψ_0.

[2] $w_0(\psi_0)$ is here the probability $\mathrm{P}(\mathcal{N}(0, \sigma^2) > 0) = 1/2$.

Example 6.8 (Test for zero contributions). Under $H_p(4) : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, U_1$ for MC15, the estimated contribution from U_1 of $\hat{\phi}_{U_1} = 0.008$ (Table 5.2) suggests that the fraction of DNA is essentially non-existent.

We can test for the removal of the unknown contributor, $\phi_{U_1} = 0$, using a single constraint. Thus, the expression (6.1) reduces to $\frac{1}{2}\,\mathrm{P}(\chi^2_1 \geq t)$. The likelihood ratio statistic for testing $H_p(3)$ against $H_p(4)$ is 0.205, which leads to a p-value of 0.325, and so it would be justifiable to remove the unknown contributor under $H_p(4)$.

Under $H_d(4) : K_1 \,\&\, K_2 \,\&\, U_1 \,\&\, U_2$, we obtained estimates $\hat{\phi}_{U_1} = \hat{\phi}_{U_2} = 0.081$ for the unknown contributors. The likelihood ratio statistic for $\phi_{U_2} = 0$, i.e. removing the smallest of the two unknown contributors, is 4, which leads to $p = 0.023$, and so we conclude that the model with two unknown contributors explains the mixture significantly better. •

Example 6.9 (Investigating shared contributors). In the examples so far, we have made the assumption that the two mixtures MC15 and MC18 share a set of unknown contributors, but evidently this may not be the case.

For the prosecution hypothesis $H_p(4) : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, U_1$ there is just a single unknown contributor shared between the mixtures. The model $H_p(5) : K_1 \,\&\, K_2 \,\&\, K_3 \,\&\, U_1 \,\&\, U_2$ has two shared unknown contributors and is thus a larger model, which allows the possibility that each of these has only contributed to one of the mixtures – this is the special case of $\phi_{1,U_2} = 0$ and $\phi_{2,U_1} = 0$. We recall that there is an order restriction only on the first mixture, so here $\phi_{2,U_1} = 0$ does not imply $\phi_{2,U_2} = 0$.

The two models of one shared versus two distinct contributors are not nested and can therefore not be compared by a likelihood ratio test. However, we can test $H_p(4)$ against $H_p(5)$ to see if the reduction to a model with one shared unknown contributor is justifiable. As there are here two constraint functions $h_j, j = 1, 2$, the p-value obtained from (6.1) will involve the computation of weights $w_j, j = 1, 2$.

However, it is possible to show from (6.1) that the p-value is at least $\frac{1}{2}\,\mathrm{P}(\chi^2_1 \geq t)$, which in our case is greater than 0.05. Thus, the assumption of a single shared unknown contributor is justifiable.

Under the defence hypothesis we would test $H_d(4)$ against $H_d(6)$, allowing the two unknowns to be different for the two mixtures. This test is expressed in terms of four constraint functions, so the associated p-value computed from (6.1) is again quite complicated. An upper bound for the p-value can be established – which gives a possibility of rejecting the hypothesis – but $\frac{1}{2}P(\chi^2_1 \geq t)$ no longer constitutes a lower bound. A simple work-around would be to test sequentially one or two constraints at the time. We do not pursue this further. •

Example 6.10 (Test for the presence of a contributor). As a thought experiment, if we want to compare $H_p(3)$ and $H_d(3)$, then these are both nested into $H_p(4)$ by setting $\phi_{U_1} = 0$ in the first case and $\phi_{K_3} = 0$ in the second case. Thus, the estimated mixture proportions for K_3 and U_1 under $H_p(4)$ can provide an idea of which of the two contributors best explain the mixture – but note, as is indeed the case, the best explanation may well be that *both* K_3 and U_1 are present.

Figure 6.8 shows the profile likelihood for ϕ_{K_3}, from which it is clear that the data point to $\phi_{K_3} \neq 0$ and thus that the profile K_3 is indeed present in the mixture. A test for reducing $H_p(4)$ to $H_d(3)$ using the halved p-value as above is firmly rejected.

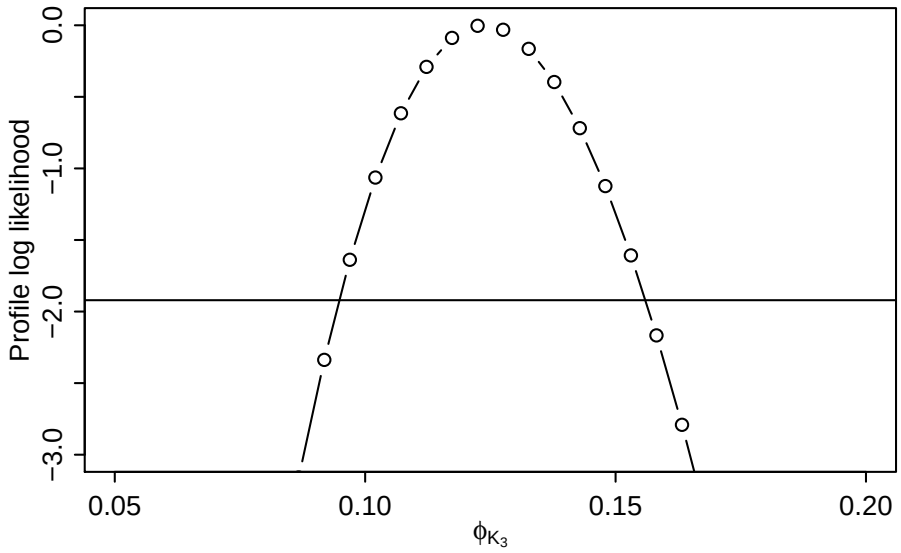


Figure 6.8: Profile log likelihood for the contribution of defendant K_3 under $H_p(4)$. Values above the horizontal line constitute an approximative 95% confidence interval.

•

The tests discussed above are based on the assumption that the model suitably explains the data, and should therefore be used in combination with other diagnostic methods. In the remainder of this chapter, we develop methods for inspecting the fit of a chosen model.

6.3 Assessing the peak height distributions

In this section we develop methodology for investigating whether a model appropriately predicts the observed peak heights.

In the discussion below we assume that the set of parameters is fixed and known, and for the notation we thus suppress the dependence on the parameters. For all examples, we use the maximum likelihood estimates under the hypothesis of interest, but note that the methods are suitable for any choice of parameters.

The distribution of a single peak height is a mixture on $\{0\} \cup [C, \infty)$; there is a point mass in zero, both because of the application of the detection threshold, and because the allele may not be present in the mixture.

One way of visualising the peak height distributions is by the use of box-and-whisker plots. In order for the boxplot to correctly reflect that the peak height distribution is a mixture, we draw only the part of the boxplot that raises above the detection threshold.

Note that the above-threshold density of a peak height may have multiple modes due to the conditional distribution of a peak height given genotypes being a mixture of gamma distributions; this behaviour is perhaps not easily visualised via boxplots.

Example 6.11 (Simulated peak heights). As a very simple tool, we may simulate from the model and thereby compare observed peak heights to their distribution under the model. We simulate under each of $H_p(4)$ and $H_d(4)$ for MC15 from the joint distribution $\Pr(\boldsymbol{Z} \mid H, \hat{\psi})$ of peak heights. The marginal variability of each single peak is indicated in Figure 6.9 by a boxplot. We have used the format for boxplots as implemented in `boxplot` in R: the box indicates the median and the upper and lower quartiles, which define the inter-quartile range (IQR), and the whiskers extend to the most extreme peak heights that are still within 1.5 IQR of the box.

In models with a large contribution from one or more unknown contributors, the variability of peak heights can be very high due to the a-priori possibility that the unknown contributors can possess any allele. This effect is seen in Figure 6.9, where under $H_d(4)$ there is more variability than under $H_p(4)$.

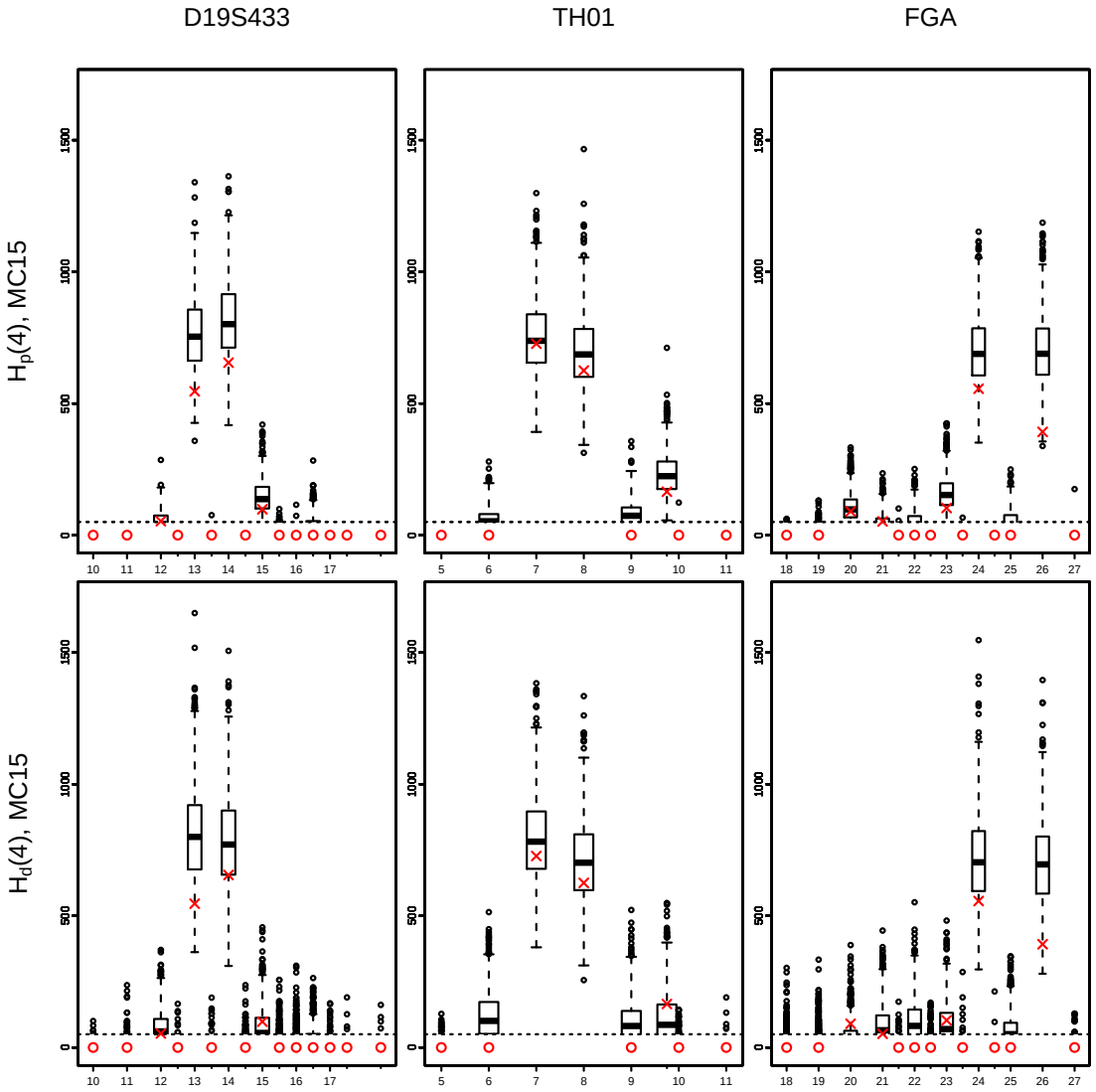


Figure 6.9: Simulated peak heights from the joint distribution of peak heights across the EPG. Observed peak heights are marked in red, with crosses indicating a peak above the threshold.

Many theoretical quantities related to the peak height distributions can be computed directly by exact methods. Below we exploit the ability to compute the cumulative distribution function and the quantiles of the distribution of peak heights.

6.3.1 The cumulative distribution function

The cumulative distribution function for a peak height Z_a can be computed by means of auxiliary variables, since

$$\mathrm{P}(Z_a \leq z) = \mathbb{E}\left\{ \mathrm{P}(Z_a \leq z \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) \right\} \tag{6.2}$$

is the expectation of a product with only one factor. To do this, we add to the model an auxiliary variable Q_a with the same parents as for O_a and with conditional probability

$$\mathrm{P}(Q_a = 1 \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) = \mathrm{P}(Z_a \leq z \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) = \begin{cases} \mathrm{P}(Z_a \leq C \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}), & z \leq C \\ \mathrm{P}(Z_a \leq z \,|\, \boldsymbol{n}_a, \boldsymbol{n}_{a+1}), & z > C. \end{cases}$$

For convenience we also introduce a set of binary variables D_a for the evaluation of both $\mathrm{P}(Z_a < C)$ and $\mathrm{P}(Z_a \geq C)$ using the definition (4.10) for auxiliary variables introduced for the representation of the qualitative model, Section 4.5.

It can be of interest to consider the distribution of the peak height in the light of other observed peaks, and not just the marginal distribution of the peak itself. For instance, we can condition on the peak heights at all other alleles to get $\mathrm{P}(Z_a \leq z \,|\, Z_b = z_b, b \neq a)$ simply through conditioning on relevant subsets of variables O_a, using Proposition 3.

Figure 6.10 shows an example of the c.d.f for a peak height, computed using one of the variables Q_a by, for a range of peak heights z, setting the probability table using z and then performing a single propagation to marginalise over $\boldsymbol{n}$.

Since we are able to evaluate the cumulative distribution function it is also possible to compute the exact quantiles of the peak height distribution. For $p <$

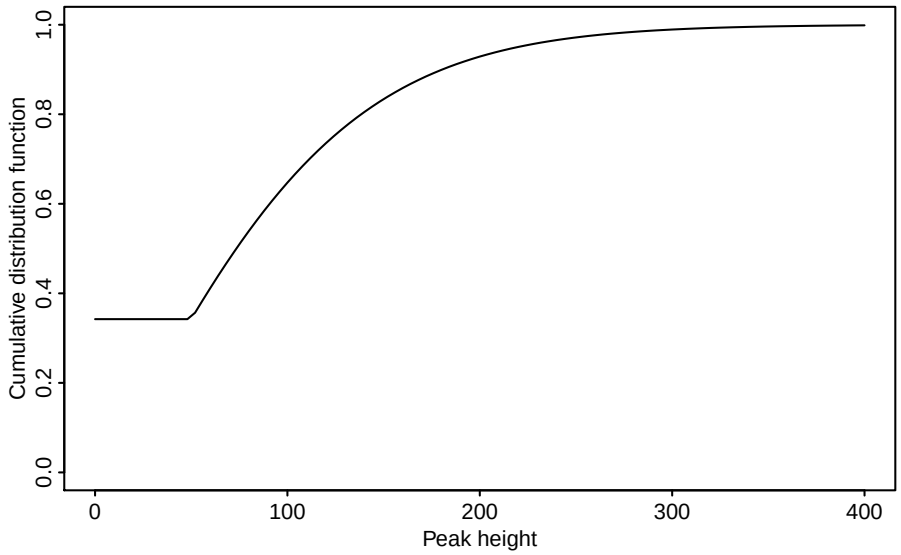


Figure 6.10: Cumulative distribution function for the peak height at marker VWA, allele 17 under $H_d(4)$. A detection threshold of $C = 50$ is applied.

$\mathrm{P}(Z_a \leq C)$, the corresponding quantile is simply zero. The c.d.f.

$$\mathrm{P}(Z_a \leq z) = \begin{cases} \mathrm{P}(Z_a \leq z), & z \geq C \\ \mathrm{P}(Z_a \leq C), & z < C \end{cases}$$

is bijective for $z \geq C$, and so for any $p \geq \mathrm{P}(Z_a \leq C)$ we may find the corresponding quantile by inverting the c.d.f; we use numerical inversion as implemented by `uniroot` in R.

Example 6.12 (Exact prediction intervals). Figures 6.11 and 6.12 visualise the conditional distribution of the peak height for an allele given the observed peak heights for all other alleles, as evaluated under $H_p(4)$ and $H_d(4)$ respectively.

Each box marks the exact 25%, 50%, and 75% quantiles of the peak height distribution. The whiskers mark the 0.5% and 99.5% quantiles, thereby giving a 99% prediction interval for the peak height. The bar under each peak then visualises how much probability mass falls below the threshold.

Due to the variability introduced by the uncertainty about the genotypes, the prediction intervals are for many of the alleles wider under $H_d(4)$ than under $H_p(4)$.

Overall, $H_p(4)$ seems to explain the peak heights better than $H_d(4)$. In Figure 6.12 we notice for instance that for marker D8S1179, the model renders the observed peaks at alleles 10 and 11 highly unlikely. In contrast, under $H_p(4)$ where these two alleles are explained by the genotype of defendant K_3, the heights of the two observed peaks are in the centre of the predictive distribution.

•

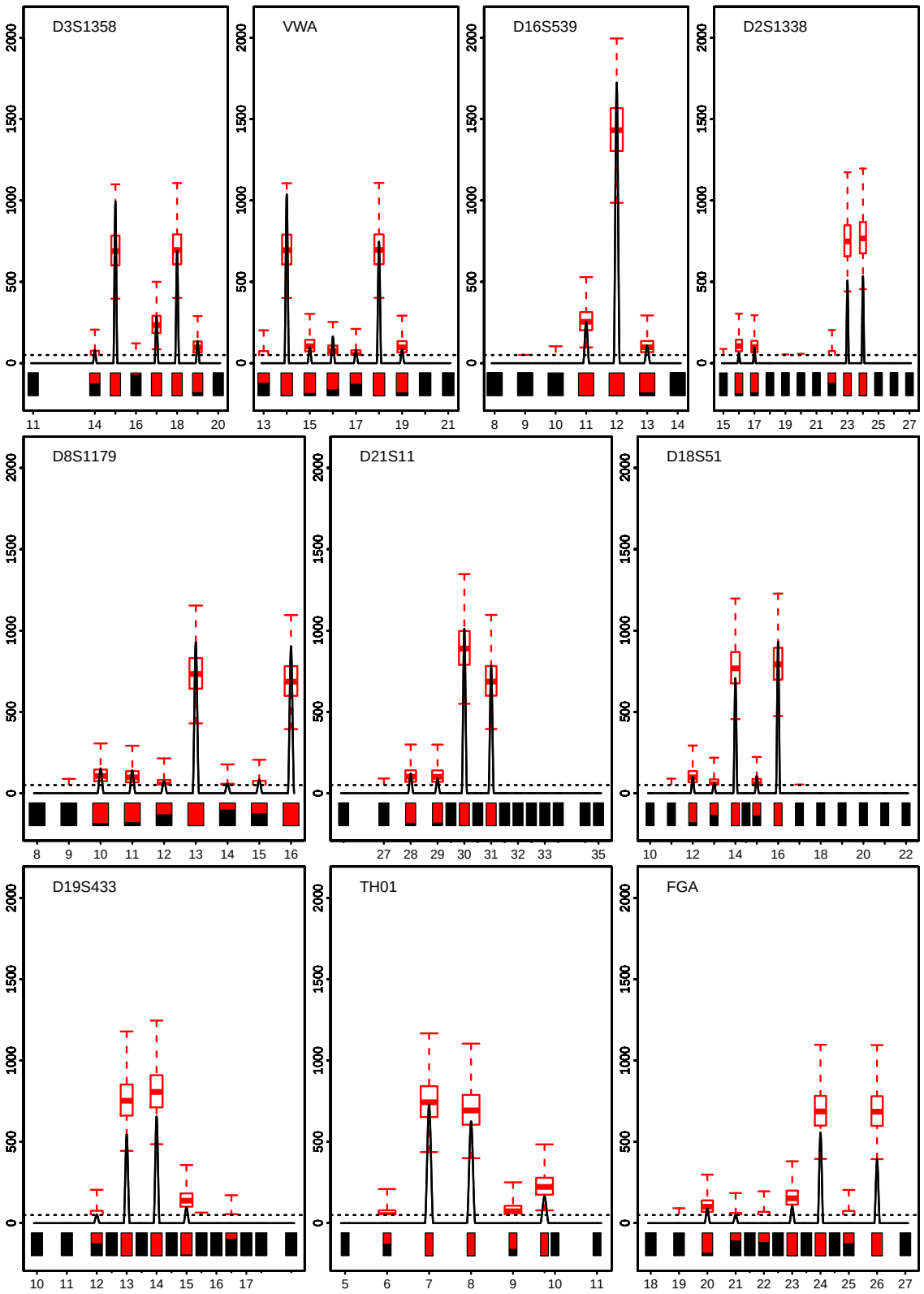


Figure 6.11: Prediction intervals for the peak heights at each allelic position under $H_p(4)$ for MC15. Intervals are based on the distribution $\mathrm{P}(Z_a \leq z_a \,|\, Z_b, b \neq a)$ of a peak height given the observed peaks heights at all other positions. Bars indicate the probability of a peak height above (red) and below (black) threshold.

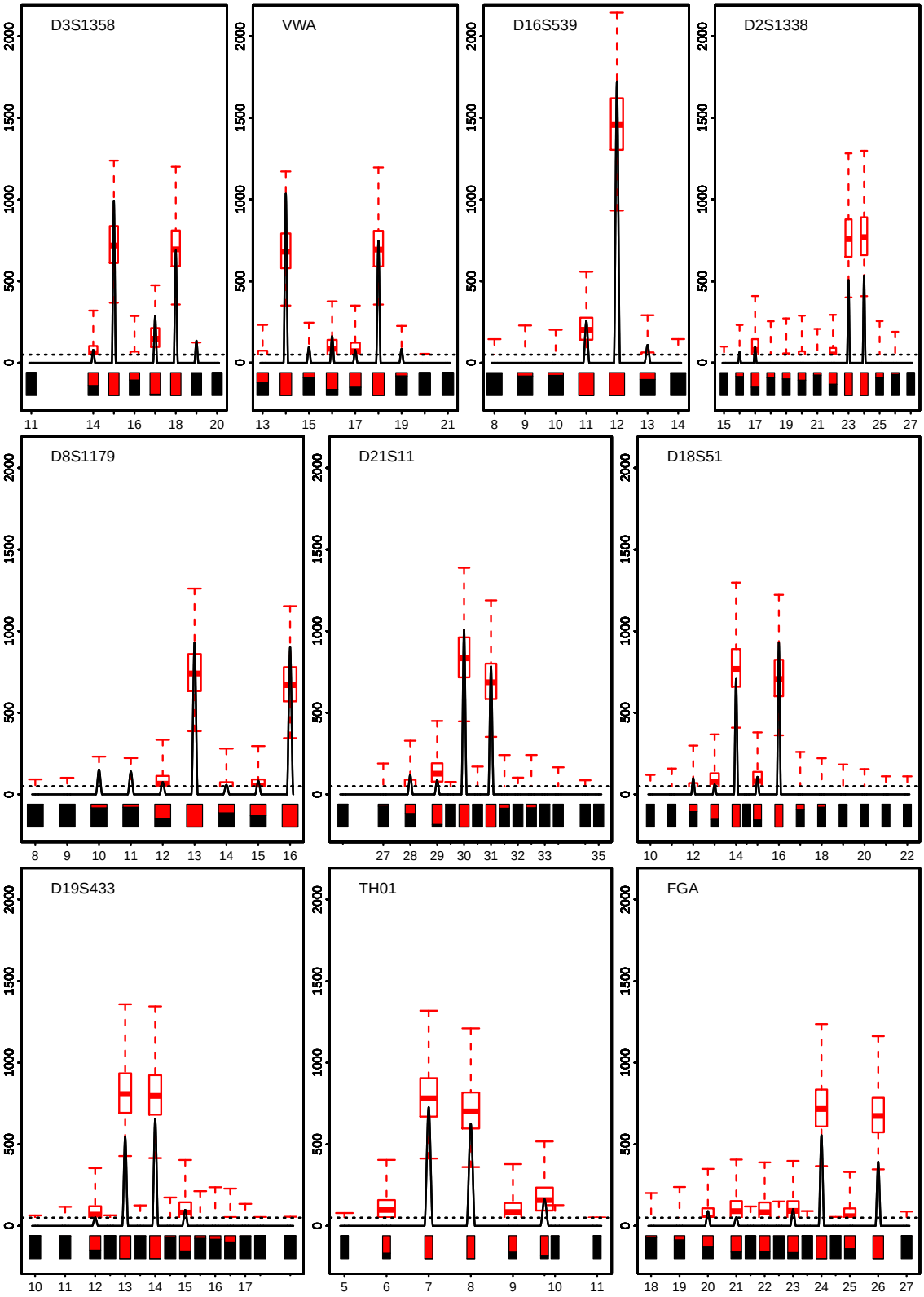


Figure 6.12: Prediction intervals for the peak heights at each allelic position under $H_d(4)$ for MC15. Intervals are based on the distribution $\mathrm{P}(Z_a \leq z_a \,|\, Z_b, b \neq a)$ of a peak height given the observed peaks heights at all other positions. Bars indicate the probability of a peak height above (red) and below (black) threshold.

From Figures 6.11 and 6.12 we may get the impression that some peak heights are systematically predicted too large or too small, and it is tempting to draw such conclusions. However, if we consider only the peaks at alleles where a peak is seen above threshold, we need to take that into account.

The methodology of the next section enables us to address questions regarding a systematic misspecification of the peak height distribution.

6.3.2 Assessing the distribution of above-threshold peaks

Example 6.13 (Inspecting peaks observed above threshold). Figure 6.13 enables a closer inspection of the peaks observed above threshold. The prediction intervals are for each allele based on the conditional distribution of the peak height given that it is above the detection threshold, and also given the observed heights at all other alleles.

Detailed inspection reveals that out of a total of 46 peaks, 25 peaks lie above the median in their respective distributions. Informally we note that, in a simple binomial test, the proportion of peaks above their median is not significantly different from a half, though we also note that a binomial test is not entirely appropriate as the observations are not independent.

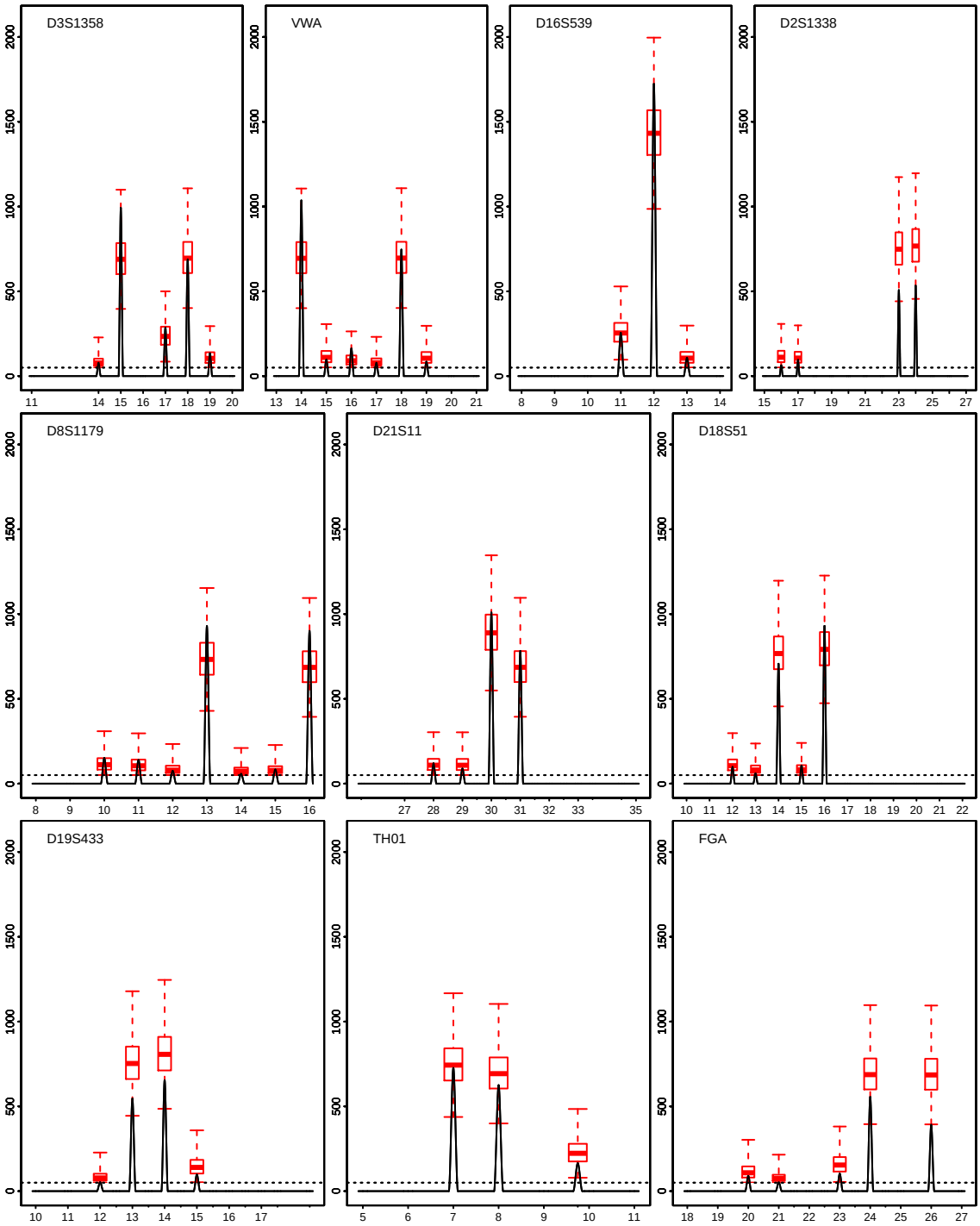


Figure 6.13: Prediction intervals based on the distribution $\mathrm{P}(Z_a \leq z_a \,|\, Z_b, b \neq a, Z_a \geq C)$ of peak heights given that a peak is observed above threshold.

•

Given that the peak is above threshold, $Z_a \geq C$, the peak height follows a continuous distribution and thus the probability transform $\mathrm{P}(Z_a \leq z_a \,|\, Z_a \geq C)$ follows a uniform distribution. We may rewrite the probability transform as

$$\mathrm{P}(Z_a \leq z \,|\, Z_a \geq C) = \frac{\mathrm{P}(Z_a \leq z) - \mathrm{P}(Z_a < C)}{\mathrm{P}(Z_a \geq C)}.$$

Thus all we need to evaluate is the distribution function in the observed value z_a and at the threshold C.

Example 6.14 (Quantile-quantile plots for MC15). In Figure 6.14, quantile-quantile plots for the conditional distribution of a peak height given observed peak heights for all other alleles are shown for both $H_p(4)$ and $H_d(4)$ using sample MC15 and the associated maximum likelihood estimates in Table 5.1.

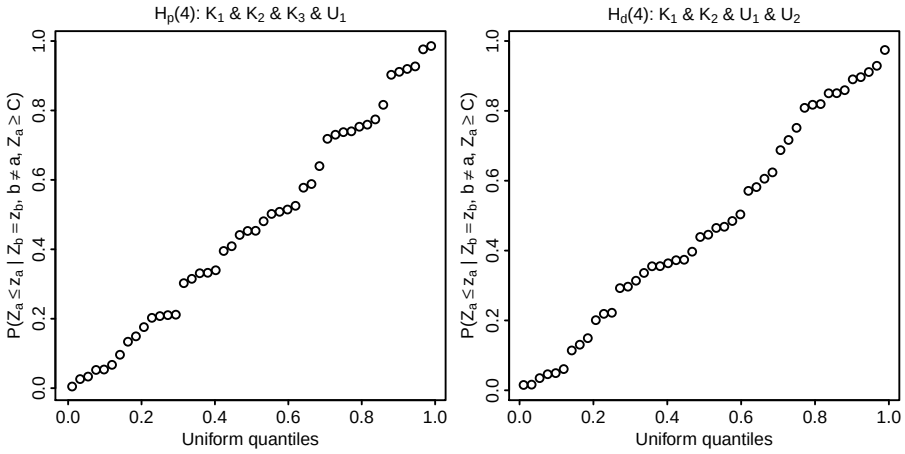


Figure 6.14: Quantile-quantile plots for the prosecution and defence hypotheses for MC15.

We note that in both diagrams the points are close to the identity line and there is no indication that the peak height distributions are inadequately modelled under either hypothesis, though the fit under $H_p(4)$ seems better than under $H_d(4)$. •

6.3.3 Checking for trends in the EPG

As only peaks above threshold contribute to the quantile-quantile plot, the plot will typically be based on few observations when analysing a single mixture. When

multiple mixtures are analysed jointly, there are more observations, giving a better opportunity to investigate whether there are systematic misspecifications of the peak height distribution; for instance, there could be signs of a dependence on dyes, markers, or the fragment length of the alleles.

Example 6.15 (Trends over mixtures). From Figure 6.15 it is clear that the peak height distributions are overall adequately modelled under both $H_p(4)$ and $H_d(4)$ for the joint analysis of MC15 and MC18.

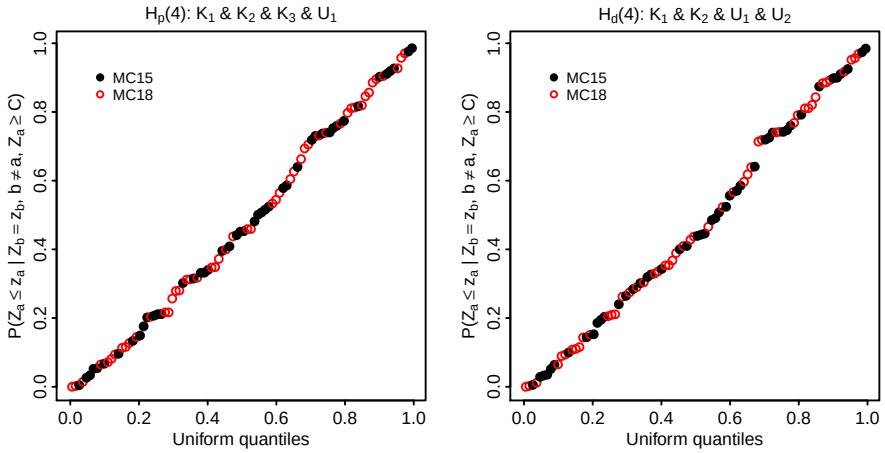


Figure 6.15: Quantile-quantile plots for the prosecution and defence hypotheses for MC15 and MC18 combined.

There is also no systematic pattern in the probability transforms corresponding to each mixture, confirming that there is no misspecification relating to the individual mixtures. •

Example 6.16 (Trends over dyes). The quantile-quantile plot for $H_p(4)$ in Figure 6.15 did not give any reason to doubt the overall fit or the fit to each mixture. However, inspection of the probability transforms grouped according to each of the three panels in the EPG does suggest a difference in peak height distribution between dyes (Figure 6.16). In particular, the yellow dye seems to exhibit lower peak heights than predicted by the model.

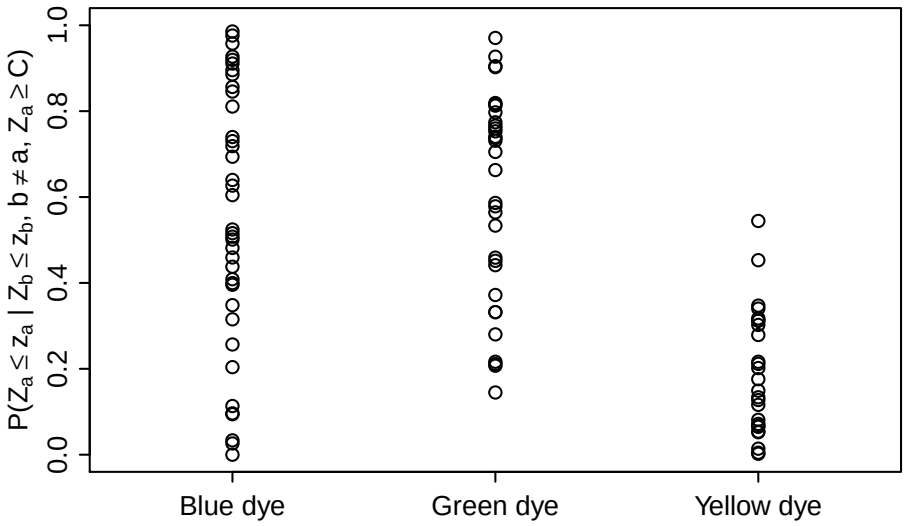


Figure 6.16: Checking for trend over dyes using probability transforms.

To investigate this we model MC15 and MC18 jointly, introducing a set of parameters for each of the three dye lanes in the EPG; we even let the mixture proportions vary freely between the dyes. To our comfort – as this is a fundamental assumption of the model – the estimated mixture proportions are highly similar.

Table 6.5: Estimates under $H_p(4)$, using one set of parameters per combination of mixture and dye.

Mixture	Dye	ρ	η	ξ	ϕ_{U_1}	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}
MC15	blue	31.2	31.1	0.057	0.015	0.802	0.067	0.116
	green	171.0	6.1	0.081	0.015	0.829	0.028	0.127
	yellow	83.8	8.4	0.065	0.023	0.829	0.032	0.117
MC18	blue	59.9	19.2	0.079	0.021	0.676	0.106	0.197
	green	111.9	10.6	0.079	0.015	0.697	0.101	0.187
	yellow	44.0	18.3	0.080	0.000	0.740	0.075	0.185

A test for reducing to a common set of mixture proportions within each mixture is strongly supported by comparing a likelihood ratio test statistic of 5.23 to the χ^2_{16}-distribution, leading to a p-value of 0.994. The 16 degrees of freedom are due to a reduction from a model with $6 \times (4-1) = 18$ parameters for the mixture

proportions and 6 stutter proportions, down to a model $2 \times (3-1) = 6$ parameters for mixture proportions and 2 stutter proportions.

Table 6.6: Estimates under $H_p(4)$, assuming a single ξ and ϕ per mixture but allowing a distinct ρ and η for each combination of dye and mixture.

Mixture	Dye	ρ	η	ξ	ϕ_{U_1}	ϕ_{K_1}	ϕ_{K_2}	ϕ_{K_3}
MC15	blue	29.4	33.0	0.075	0.013	0.826	0.038	0.124
	green	173.6	6.0	0.075	0.013	0.826	0.038	0.124
	yellow	75.6	9.3	0.075	0.013	0.826	0.038	0.124
MC18	blue	62.3	18.5	0.080	0.017	0.699	0.093	0.191
	green	102.1	11.6	0.080	0.017	0.699	0.093	0.191
	yellow	36.8	21.9	0.080	0.017	0.699	0.093	0.191

Further reduction to the model $H_p(4)$, which assumes a common ρ and η for all markers in a mixture, is firmly rejected: the likelihood ratio test statistic of 62.8 compared to χ^2_8 leads to a p-value of 10^{-10}.

It may seem that ρ exhibits more similarity across mixtures than across dyes, and it can well be an effect of some alleles amplifying more than others. In the further investigation on such matters, it is important to bear in mind the interpretation of ρ as being proportional to the total amount of DNA in the mixture. Thus, further structural assumptions about ρ would naturally involve some mixture-specific component.

A common assumption is that peak heights depend on the length of the allele. Figure 6.17 does show some decline for longer allele lengths, although the decline seems to be dependent on the dye as well. It would seem from Figure 6.17 that a stronger trend lies in the difference between dyes.

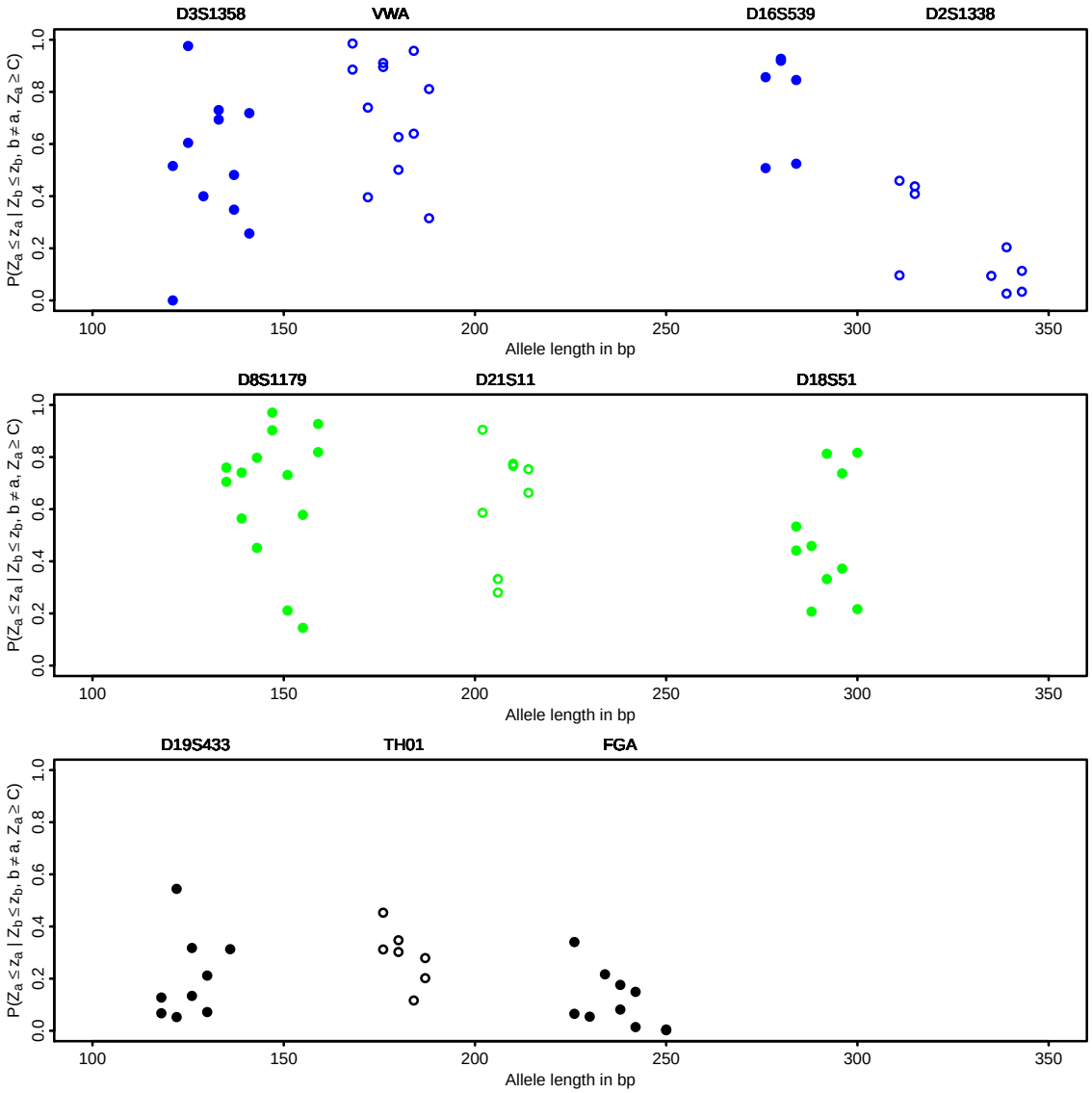


Figure 6.17: Checking for trends in the EPG using probability transforms of the peak heights above threshold. The non-overlapping groups of points corresponding to each marker are highlighted by the use of two different plot symbols.

Marker D2S1338 to the right in the blue dye lane exhibits noticeably lower peak heights than predicted under the model, and seemingly more than simply due to a trend over allele length; we have no good explanation for this.

A similar plot for the model using different parameters for each dye shows that it does indeed correct the overall fit, but the trends within each panel somewhat persists; we omit these plots.

A further investigation of the trend over allele lengths – particularly on larger sets of experimental data – would be useful for the assessment of the general suitability of the gamma model in its current form as well as extensions to allow for degradation. However, the suitability of the model to a particular case should still be assessed. •

6.4 Prequential monitoring of peak presence

Our next diagnostic method is concerned with whether the model correctly predicts absence and presence of peaks in the EPG. The method is based on the prequential theory of Dawid (1984) with so-called prequential monitors (Seillier-Moiseiwitsch and Dawid, 1993).

Using any ordering of alleles in the EPG, we consider the set of alleles across all markers. Let p_a be the probability that a peak has been seen for allele a given the peak heights observed on all preceding alleles,

$$p_a = \mathrm{P}(Z_a \geq C \,|\, z_i, i < a) = \mathrm{P}(D_a = 0 \,|\, z_i, i < a).$$

This probability can be obtained after a propagation of likelihood evidence for the nodes $(O_a, a < i)$ as described in Section 4.3. For each allele a, we then consider the logarithmic score

$$Y_a = \begin{cases} -\log p_a, & z_a \geq C \\ -\log(1 - p_a), & z_a < C \end{cases}$$

so that Y_a is always non-negative and higher values of Y_a represent a large penalty for assigning a small probability (p_a or $1 - p_a$) to the event that is actually observed.

The cumulative logarithmic score, adjusted for incremental expectations,

$$M_a = \sum_{i=1}^{a} \{Y_i - \mathbb{E}(Y_i \,|\, Z_b, b < i)\}$$

is a martingale with respect to the sequence of peak heights. As

$$\mathbb{V}(M_a - M_{a-1} \mid Z_b, b < a) = \mathbb{V}(Y_a \mid Z_b, b < a),$$

the distribution of the normalised cumulative score

$$\frac{\sum_{i=1}^{a} Y_i - \sum_{i=1}^{a} \mathbb{E}(Y_i \mid Z_b, b < i)}{\sqrt{\sum_{i=1}^{a} \mathbb{V}(Y_i \mid Z_b, b < i)}}$$

approaches a standard normal distribution as the denominator becomes infinitely large (Seillier-Moiseiwitsch and Dawid, 1993). Thus, for $q_{1-\alpha}$ denoting the $1-\alpha$ quantile of the standard normal distribution,

$$q_{1-\alpha} \sqrt{\sum_{i=1}^{a} \mathbb{V}(Y_i \mid Z_b, b < i)}$$

is an approximate pointwise $1-\alpha$ upper predictive limit for the cumulative score at allele a.

The cumulative score can easily be calculated using that if $p_a \in \{0, 1\}$ we have $Y_a = 0$ and otherwise

$$\begin{aligned}
\mathbb{E}(Y_a \mid Z_b, b < a) &= -p_a \log p_a - (1 - p_a) \log(1 - p_a), \\
\mathbb{V}(Y_a \mid Z_b, b < a) &= p_a(1 - p_a) \left\{\log p_a - \log(1 - p_a)\right\}^2.
\end{aligned}$$

The prequential monitor is a graph showing the development in the cumulative score as we consider more and more alleles; an example is seen in Figure 6.18 below.

A negative jump in the score means that we have observed what the model predicts as most likely, whereas a positive jump means that we have observed something unlikely, namely the opposite of what is most likely according to the model. If it is equally likely for a peak to fall above and below the threshold, or there is only one possible outcome – i.e. if $p_a \in \{0, 0.5, 1\}$ – there is no jump. The size of an upward jump indicates the level of disagreement between model and observations. Investigation of the upward jumps may reveal whether they are observation of rare alleles, a misfit of the dropout model, or even to recording errors in the data. Note

that the order in which the alleles are considered does matter for the appearance of the monitor, but the value at the final observation will remain the same.

Example 6.17 (Prequential monitors). Prequential monitor plots of the prosecution and defence hypotheses for MC15 and four contributors are displayed in Figure 6.18. Overall, both hypotheses adequately predicts the presence and absence of peaks; the end-point of the prequential monitor is in both cases well below the predictive limits.

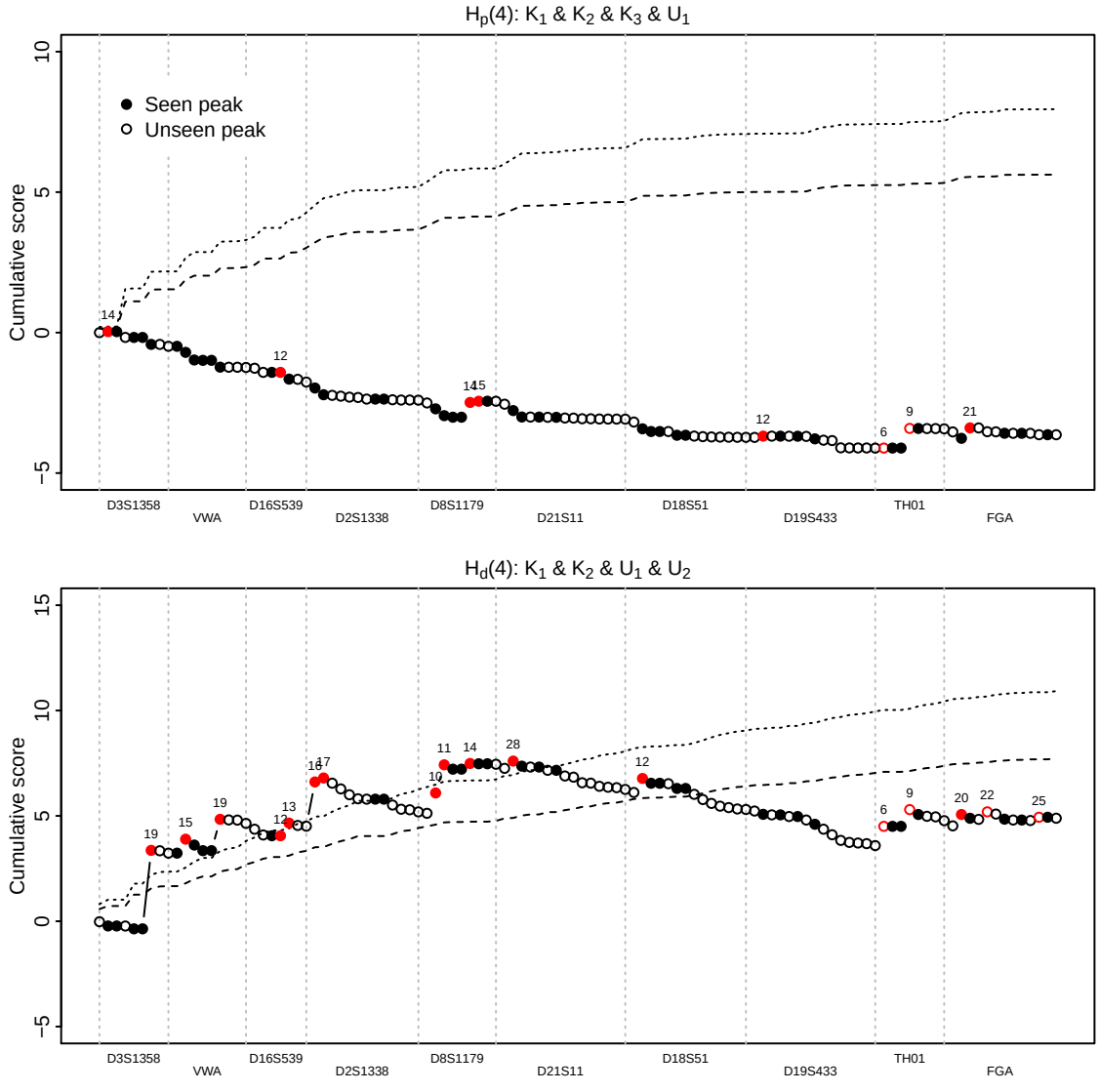


Figure 6.18: Prequential monitor plots of the prosecution and defence hypotheses for MC15. Upward jumps are marked in red. Upper 95% (dashed) and 99% (dotted) pointwise predictive limits based on the approximating normal distribution are indicated.

By closer inspection, we notice that most of the upward jumps are due to peaks that can solely be attributed to the unknown contributors. We believe that this is due to the nature of the distributional assumption on unknown contributors; for any allele there would only be a small probability that an unknown contributor possesses exactly that allele. The exact interpretation of the distribution for the genotypes of an unknown contributor is generally tricky; it reflects more a subjective belief about the genotypes than a model. A further few jumps toward the end of the plot are due to the dropped out alleles of individual K_2.

The tests in Example 6.8 indicated that an unknown contributor could be removed from $H_p(4)$, and this is underlined by the prequential monitor virtually being unchanged (plot not shown). Under the defence, however, the hypothesis $H_d(4)$ was significantly better than $H_d(3)$. The prequential monitor for $H_d(3)$ in Figure 6.19 crosses the upper limits and stays there until the end of the plot, supporting the conclusion that this hypothesis may not adequately describe the pattern of observed peaks.

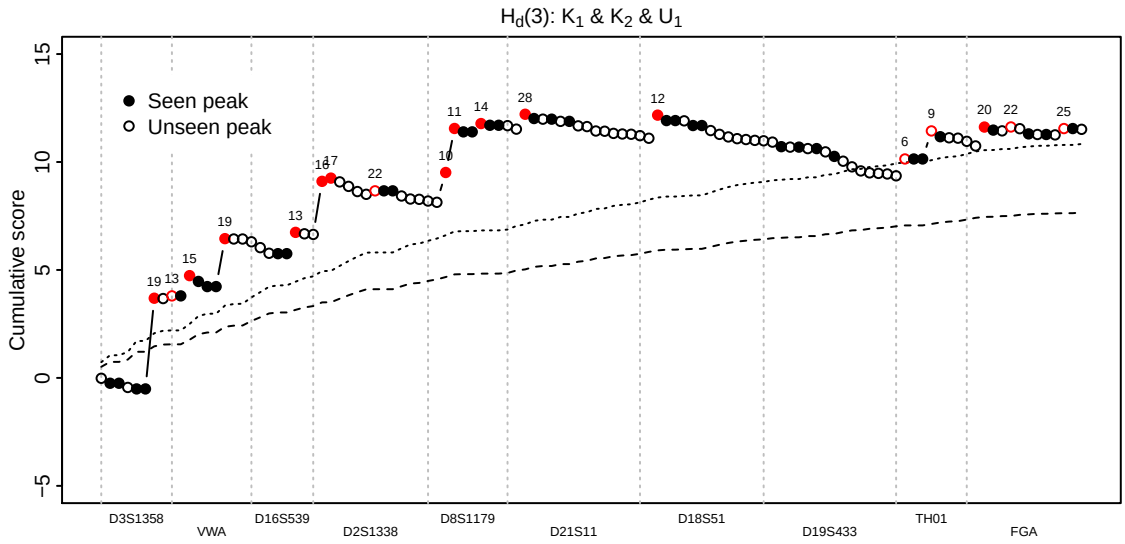


Figure 6.19: Prequential monitor for $H_d(3)$: K_1 & K_2 & U_1 for MC15 with upper 95% (dashed) and 99% (dotted) pointwise predictive limits based on the approximating normal distribution. Upward jumps are marked in red.

The diagnostic tools presented in this and the previous section are all based on the assumed distribution of peak heights and address whether there are signs of misspecification. Ideally we would also directly challenge the proposed distribution for the unknown contributors under a hypothesis, since a misspecification of the genotype distribution would translate to misspecification of the distribution of peak heights. This would also be an effect of an inadequate conditional peak height distribution given genotypes.

Chapter 7

The DNAmixtures package

The R-package DNAmixtures (Graversen, 2014) implements Model 2.3 for the analysis of one or more mixtures and makes available the statistical methodology of Chapters 5 and 6. All analyses in this thesis as well as those in Cowell et al. (2015) and Graversen and Lauritzen (2014) have been carried out using DNAmixtures. The package can be found at the project homepage

`http://dnamixtures.r-forge.r-project.org/`

along with a guide to installation. Once installed, the package can be loaded as

```
> library(DNAmixtures)
```

The implementation of DNAmixtures is based on computation with auxiliary variables, presented in Section 3.2, combined with the efficient network representation of the model described in Chapter 4. It relies on the standard machinery for computations in Bayesian network models provided by HUGIN and does so through the interface provided in the R-package RHugin (Konis and Hugin Expert A/S, 2014).

The DNAmixtures package has no graphical user interface and requires a basic familiarity with R. The aim has been to provide an intuitive and userfriendly interface without compromising general flexibility.

Below we give a brief overview of the functionality of the package with a few examples of the core data structures and functions. For details and more elaborate examples, we refer to the manual pages of DNAmixtures.

7.1 The `DNAmixture`

The `DNAmixture` provides a full representation of the joint model $\Pr(\boldsymbol{Z}, \boldsymbol{n} \mid H, \psi)$, Model 2.3, of DNA profiles $\boldsymbol{n}$ for the contributors to one or multiple DNA mixtures and the peak heights $\boldsymbol{Z}$ in their corresponding EPGs. To specify this model we need

- The observed peak heights and the detection threshold C used for each of the mixtures included in the model.
- The hypothesis H specifying the set of contributors in terms of
  - The number k of contributors
  - The set of known DNA profiles
  - The allele frequencies, which are then used in the standard model for the set of unknown contributors (Section 2.1).

The set of model parameters ψ is typically given as an argument to functions acting on a `DNAmixture` object.

Peak height data. Observed peak heights are specified in a `data.frame` containing at least the variables `marker`, `allele`, and `height`. The marker names given in `marker` can be chosen arbitrarily but should be used consistently within the specification of a `DNAmixture`. An exception to this is the marker Amelogenin, which should always be named `"AMEL"`. The variable `allele` should be numeric and contain the repeat numbers.

The peak heights for MC15 are included as example data in the package:

```
> data(MC15)
> MC15[MC15$marker == "TH01",]

   marker allele height K1 K2 K3
35   TH01    7.0    727  1  0  0
36   TH01    8.0    625  1  0  0
37   TH01    9.0      0  0  2  0
38   TH01    9.3    165  0  0  2
```

Known contributors. The profiles of known contributors are specified as allele counts and can be specified together with the peak heights, as for MC15 above, or in a separate `data.frame` via the optional argument `reference.profiles` for `DNAmixture`.

There should be one variable per known contributor, each containing the allele count for all combinations of variables `marker` and `allele`. The names for the known contributors can be chosen freely.

Allele frequencies. The allele frequencies are also specified in a `data.frame`, this one containing variables `marker`, `allele`, and `frequency`. The range of alleles in this dataset determines the range of alleles used in the model. Any alleles found in the peak height data and known contributor data are also included, but assigned a frequency `NA` if no allele frequency exists for this allele. Alleles with frequency zero are allowed.

Note that in the current implementation the stutter proportion for an allele a is set to zero if $a - 1$ is not within the range of alleles. The US-Caucasian allele frequencies are included in the package:

```
> data(USCaucasian)
> USCaucasian[USCaucasian$marker == "TH01",]

   marker allele   frequency
22   TH01    5.0 0.001659967
23   TH01    6.0 0.231785364
24   TH01    7.0 0.190396192
25   TH01    8.0 0.084438311
```

```
26   TH01    9.0 0.114237715
27   TH01    9.3 0.367542649
28   TH01   10.0 0.008279834
29   TH01   11.0 0.001659967
```

As an example, let us create a `DNAmixture` for the model $H_p(4) : K_1 \& K_2 \& K_3 \& U_1$ for the single DNA mixture MC15.

```
> mix15p4 <- DNAmixture(
      list(MC15),             # Peak heights and known profiles
      C = list(50),           # Detection threshold
      k = 4,                  # Number of contributors
      K = c("K1", "K2", "K3"),# Names of known contributors
      database = USCaucasian  # Allele frequencies
      )
> mix15p4

A DNA mixture model with 4 contributors.

Known: K1 K2 K3
Unknown: U1

Mixtures included: list(MC15)
Detection threshold(s): 50
```

For a joint model of multiple DNA mixtures, these are simply included in the list specifying peak height data. The mixtures can have all, some, or no markers in common. They may also have individual detection thresholds, and these are specified in the list `C`.

To build the joint model $H_p(4)$ for MC15 and MC18, we firstly load the dataset `MC18`, which is included in the package.

```
> data(MC18)
> mix1518p4 <- DNAmixture(
      list(MC15, MC18),       # Peak heights and known profiles
      C = list(50, 50),       # Detection thresholds
      k = 4,                  # Number of contributors
      K = c("K1", "K2", "K3"),# Names of known contributors
      database = USCaucasian  # Allele frequencies
      )
> mix1518p4
```

```
A DNA mixture model with 4 contributors.

Known: K1 K2 K3
Unknown: U1

Mixtures included: list(MC15, MC18)
Detection threshold(s): 50 50
```

Parameters

The parameter ψ is represented by an object of class `mixpar`, which is specified by giving for each of the four model parameters a list containing one parameter per mixture in the model.

A parameter for $H_p(4)$ for MC15 would be specified as:

```
> psi <- mixpar(rho = list(25),
                eta = list(20),
                xi = list(0.07),
                phi = list(
                    c(K1 = 0.7, K2 = 0.05,
                      K3 = 0.1, U1 = 0.05)))
```

For each mixture, the vector `phi` of mixture proportions should be named using the labels of the contributors under the hypothesis. The names of known contributors should correspond to those specified for the `DNAmixture`. The names of unknown contributors are always named `U1`, ..., `Up` in a model with `p` unknown contributors.

Note that although not meaningful in terms of the model, mixture proportions are allowed not to sum to 1. The order of the proportions is unimportant.

The `mixpar` object is simply a two-way array of lists, where the first dimension (rows) corresponds to the DNA mixtures analysed, and the second dimension

(columns) corresponds to the parameters (ρ, η, ξ, ϕ) as

$$\begin{array}{c} \\ 1 \\ \vdots \\ R \end{array} \begin{array}{c} \begin{array}{cccc} \rho & \eta & \xi & \phi \end{array} \\ \begin{pmatrix} \rho_1 & \eta_1 & \xi_1 & \phi_{11}, \ldots, \phi_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_R & \eta_R & \xi_R & \phi_{R1}, \ldots, \phi_{Rk} \end{pmatrix} \end{array}.$$

Marker-dependent parameters. In the current implementation of DNAmixtures, parameters (ρ, η, ξ, ϕ) are the same across markers and alleles. An interesting special case of modelling multiple mixtures, or EPGs, jointly is that of modelling a single mixture, which is split into subsets of markers. This enables an analysis in which parameters and the detection threshold are specific to each of the subsets of markers. For such tricks to be meaningful, one would typcially want to use common mixture proportions for all markers relating to the same DNA sample; see for instance Example 6.16.

7.1.1 The Bayesian network representation

If a hypothesis includes unknown contributors, the implementation relies on Bayesian networks and the `DNAmixture` then contains a full network represenation of the model corresponding to that discussed in Chapter 4.1.

The networks are stored as part of the `DNAmixture` in a list indexed by the marker names; this is accessed as `domains`. For instance, we can obtain the network corresponding to marker TH01 by

```
> mix15p4$domains$TH01

  A Hugin domain: there are 40 nodes and 64 edges
```

A network for a marker m with alleles $a = 1, \ldots, A_m$ contains the following variables. For each of the unknown contributors $i = 1, \ldots, p$, there is a network of the structure in Figure 4.1 containing the allele counts `n_i_a` (n_{ia}) as well as the

partial allele counts `S_i_a` (S_{ia}). Further, there are three sets of auxiliary variables `O_e_a` (O_{ea}), `Q_e_a` (Q_{ea}), and `D_e_a` (D_{ea}) for each allele `a` and each EPG e that includes the specific marker.

Preparation of the networks

When a `DNAmixture` is created, networks are by default created, compiled and compressed. During compilation, the network is triangulated using the elimination order (4.9), so that the "split" junction tree representation of Section 4.4 is used.

Note that as HUGIN networks (of class `hugin.domain`) are merely pointers in R they cannot be saved in a workspace; thus, any `DNAmixture` with unknown contributors will have to be created anew for each R session. In the initiation of a `DNAmixture` there is an option to save or load existing networks in order to avoid re-building the networks. The size of the networks can, as was seen in Figure 4.8, be quite significant before compression, so creating the networks – perhaps even just one at the time – on a computer with more memory can be beneficial for particularly complex models.

Setting tables for auxiliary variables

The functions `setCPT.O`, `setCPT.D`, `setCPT.Q` are used internally for setting the conditional probability tables for auxiliary variables O_a, D_a, and Q_a. The wrapper `setCPT` sets the conditional probability tables for all auxiliary variables, and returns the scaling factors computed by `setCPT.O`; these are needed in the various computations by auxiliary variables, following the framework of Chapter 3.2.

The effect of calling `setCPT` is that the bayesian networks become a full representation of the model $\Pr(\boldsymbol{Z}, \boldsymbol{n} \mid H, \psi)$.

Ensuring a correct network representation

A difficulty in implementing DNAmixtures has been to keep track of the state of the Bayesian networks. The problem is that as a `hugin.domain` is a pointer, all copies of a `DNAmixture` will point to the same networks. Thus, the networks in a `DNAmixture` object are potentially changed globally by any function that performs computations on the networks, even when the computations are performed using a local copy of the `DNAmixture`. This is in contrast to the paradigm of R that objects have to be specifically overwritten for any change to take effect.

We have adopted the convention that features relating to the distribution represented by the networks are not maintained in the `DNAmixture` object, and functions using the networks for computations may therefore leave these in an unspecified state.

To ensure their correctness, functions that use the networks for computations – and therefore need them to be in a particular state – will as a default reset the networks by retracting any evidence. However, some functions allow an optional argument, `initialize`, which can be set to `FALSE` for instance when user-specified propagated evidence should be retained in the network.

7.1.2 Conditional distributions

The `DNAmixture` readily provides a representation of the conditional distribution $\Pr(\boldsymbol{n} \mid \boldsymbol{Z}, H, \psi)$, i.e. the posterior distribution of the DNA profiles (and other network variables). This is obtained through the function `setPeakInfo`, which propagates the information about peak heights in the networks of the `DNAmixture`.

Note that the evidence may need to be set again after calls to other functions operating on the `DNAmixture`, since these may alter the state of the networks.

```
> setPeakInfo(mix15p4, psi)
```

If only the discrete information about peak presence should be taken into account, this can be specified in an optional argument by `presence.only = TRUE`. To reset the `DNAmixture` to represent $\Pr(\boldsymbol{Z}, \boldsymbol{n} \mid H, \psi)$, i.e. no longer condition on the observed peak information, use

```
> removePeakInfo(mix15p4)
```

7.1.3 The likelihood function

The log-likelihod function can be obtained and evaluated as

```
> loglik <- logL(mix15p4)
> loglik(psi)

[1] -420.5525
```

Thus, `logL` returns a function – here we have named it `loglik` – for calculating the log likelihood, and this takes a `mixpar` parameter as its argument.

Importantly, by specifying `presence.only = TRUE` we obtain instead the likelihood function (2.14) based only on the presence and absence of peaks.

```
> logL(mix15p4, presence.only = TRUE)(psi)

[1] -27.39619
```

There is no specialised function for computing the WoE or the likelihood ratio, but the value of the likelihood or the maximised likelihood can be obtained under each hypothesis of interest and then compared in a suitable fashion. For the purpose of computing the WoE, where typically $\log_{10} L$ is used, note that this can be obtained from the natural log by dividing by $\log(10)$.

When there are unknown contributors in the hypothesis of the `DNAmixture`, the likelihood function is evaluated using the auxiliary variables `O_e_a`. For models involving known contributors only, the likelihood function is simply a product of

probability density functions and cumulative distribution functions for the gamma distribution.

7.2 Maximum likelihood estimation

The model parameters may be estimated by maximum likelihood via the function `mixML`. From this, both the estimated parameters and the value of the maximised likelihood can be obtained.

Here we maximise the likelihood function using `psi` as the starting point for the numerical maximisation.

```
> ml15p4 <- mixML(mix15p4, psi)
> ml15p4$mle

      rho     eta       xi     phi.U1   phi.K1    phi.K2   phi.K3
1   34.24   26.67   0.0737   0.008433   0.8205   0.04735   0.1237
```

The `mle` element returned by `mixML` is a `mixpar` and can be passed along to functions requiring a model parameter.

The maximised likelihood obtained from `mixML` is naturally the same as evaluating the log-likelihood function directly in the found MLE,

```
> ml15p4$lik

[1] -271.8025

> logL(mix15p4)(ml15p4$mle)

[1] -271.8025
```

7.2.1 Imposing constraints on the parameters

The maximisation of the log-likelihood function relies on the function `solnp` in package Rsolnp (Ghalanos and Theussl, 2012; Ye, 1987) and allows the user to impose non-linear equality constraints by specifying a function of the parameters and a value

that the function should attain. In Cowell et al. (2015), a common ξ and η was used for MC15 and MC18, which can be specified via the function

```
> eq.eta.xi <- function(q){
    c(q[[1,"xi"]]-q[[2,"xi"]], q[[1,"eta"]]-q[[2,"eta"]])
  }
```

The constraint can now be phrased as `eq.eta.xi(q) == c(0,0)`, and we may impose this for the maximisation in the model $H_p(4)$ as follows.

```
> p <- mixpar(rho = list(32, 37),
              eta = list(27, 27),
              xi = list(0.08, 0.08),
              phi = list(
                  c(K1 = 0.8, K2 = 0.05, K3 = 0.1,
                    U1 = 0.05),
                  c(K1 = 0.7, K2 = 0.09, K3 = 0.9,
                    U1 = 0.01)))
> ml1518p4 <- mixML(mix1518p4, p,
                    constraints = eq.eta.xi, val = c(0,0))
> ml1518p4$mle

      rho     eta        xi     phi.U1   phi.K1    phi.K2   phi.K3
1   32.66   27.99   0.07935   0.006043   0.8218   0.04764   0.1245
2   37.68   27.99   0.07935   0.012269   0.7045   0.08989   0.1933
```

Although specifying a constraint that some parameters are equal would effectively reduce the dimension of the parameter space, `solnp` does not seem to make this reduction. Therefore, specifying such constraints by instead making a customised likelihood function will speed up computations; this is done for the special case of equal mixture proportions across a set of mixtures as specified by `phi.eq = TRUE`. The current implementation does not allow the user to specify a likelihood function.

In the parameter space for the maximisation, proportions for unknown contributors are as default ordered by contributions to the first mixture such that

$$\phi_{1,\texttt{U1}} \geq \cdots \geq \phi_{1,\texttt{Up}}$$

for a mixture with `p` unknown contributors. Since `solnp` does not allow redundant constraints, it may be necessary to remove this ordering when further constraints are imposed on the proportions, and this can be done by setting `order.unknowns = FALSE`.

Profile likelihood

The specification of constraints can be used in computing profile likelihoods. For instance, the profile likelihood for K_3 in Figure 6.8 was computed along the lines of

```
> proflik <- function(x){
      mixML(mix15p4, ml15p4$mle,
            constraints = function(p)p[[1, "phi"]]["K3"]],
            val = x)$lik
  }
> proflik <- Vectorize(proflik)
```

The last call to `Vectorize` ensures that the function `proflik` can be given a vector of values of ϕ_{K_3} as its argument. Due to the flexibility in specifying constraints, we can easily compute the profile likelihood for more complicated parameter functions, for instance the difference between unknown contributors.

7.2.2 Asymptotic variance

The asymptotic variance for the MLE is estimated as described in Section 6.1.2. The Hessian of the negative log likelihood is found by numerical differentiation using `hessian` in the numDeriv package. For each of the four model parameters it is possible to specify them as fixed, equal or different across traces analysed.

The argument `npars` is a list specifying for each parameter whether it is fixed (0), common across mixture (1), or free (the number of mixtures). It is entirely the responsibility of the user to ensure a consistency between these specified dimensions and the model, under which the specified MLE is computed.

The function `varEst` returns a multitude of covariance matrices; for a list of estimates and their asymptotic standard errors, we use the `summary` function.

```
> var15p4 <- varEst(mix15p4, ml15p4$mle, npars =
                    list(rho = 1, eta = 1, xi = 1, phi = 1))
> summary(var15p4)

            Estimate    StdErr
rho.1      34.238776   7.13122
eta.1      26.667544   5.61818
xi.1        0.073703   0.01441
phi.U1.1    0.008433   0.01852
phi.K1.1    0.820516   0.02014
phi.K2.1    0.047348   0.01361
phi.K3.1    0.123703   0.01532
```

For $H_p(4)$ where ξ and η are assumed common for the two mixtures, we may compute the asymptotic variance matrix as

```
> var1518p <- varEst(mix1518p4, ml1518p4$mle, npars =
                     list(rho = 2, eta = 1, xi = 1, phi = 2))
> summary(var1518p, transform = TRUE)

              Estimate     StdErr
mu.1        914.218369   35.69127
mu.2       1054.689998   38.30652
sigma.1       0.174982    0.01286
sigma.2       0.162913    0.01192
xi.1          0.079346    0.01069
xi.2          0.079346    0.01069
phi.U1.1      0.006043    0.01848
phi.K1.1      0.821775    0.02033
phi.K2.1      0.047640    0.01386
phi.K3.1      0.124541    0.01567
phi.U1.2      0.012269    0.01743
phi.K1.2      0.704538    0.02150
phi.K2.2      0.089893    0.01562
phi.K3.2      0.193300    0.01772
```

In the summary above, we have specifically requested that estimates and their standard errors are returned in the parameterisation using μ and σ rather than ρ and η.

7.3 Prediction

7.3.1 Fitted probabilities

The function `predict` gives a `data.frame` containing for each allele in the EPG(s) the probabilities of a `seen` resp. `unseen` of a peak as well as the probabilities of obtaining a `larger` resp. `smaller` peak than the one observed.

```
> pred <- predict(mix15p4, ml15p4$mle,
                  dist = "conditional", marker = "TH01")
> head(pred)

$TH01
  allele         seen       unseen    larger   smaller trace height
1    5.0 6.498108e-04 9.993502e-01 1.0000000 0.0000000     1      0
2    6.0 5.028765e-01 4.971235e-01 1.0000000 0.0000000     1      0
3    7.0 1.000000e+00 1.445432e-23 0.5466311 0.4533689     1    727
4    8.0 1.000000e+00 2.289263e-21 0.6975336 0.3024664     1    625
5    9.0 7.248061e-01 2.751939e-01 1.0000000 0.0000000     1      0
6   10.0 4.036336e-04 9.995964e-01 1.0000000 0.0000000     1      0
7   11.0 8.106189e-05 9.999189e-01 1.0000000 0.0000000     1      0
8    9.3 9.997454e-01 2.546317e-04 0.7971818 0.2028182     1    165
  marker
1   TH01
2   TH01
3   TH01
4   TH01
5   TH01
6   TH01
7   TH01
8   TH01
```

There are three readily available choices `dist` of peak height distributions, under which we may evaluate the probabilities.

`"joint"`[1] No conditioning on observed peak heights, $\Pr(Z_a \mid H, \psi)$.

`"conditional"` Conditional distribution given the heights for all peaks, except the one under consideration, $\Pr(Z_a \mid Z_b, b \neq a; H, \psi)$.

`"prequential"` Conditional distribution given the heights for all peaks "before" the peak under consideration, $\Pr(Z_a \mid Z_b, b < a; H, \psi)$. For models with mul-

[1]The name is misleading and should be changed in future versions of DNAmixtures; indeed, the distribution in question is the marginal distribution of a single peak height.

tiple mixtures, the order can be controlled through the optional argument `by.allele` grouping observations according to either alleles or mixtures.

The probabilities are computed using the auxiliary variables Q_a and D_a as described in Section 6.3.

7.3.2 Highest probability sets of genotypes

The methodology of Section 5.3 for the prediction of genotypes is implemented by the function `map.genotypes`, which returns the best configurations of allele counts according to the distribution currently represented by the `DNAmixture`. The typical distribution of interest is the posterior distribution of genotypes given the observed peak height information.

There are three choices of subsets of alleles – `all`, `seen`, and `unseen` – allowing the user to summarize the distribution of genotypes in a few ways as discussed in Section 5.3. Further, the prediction for a subset of the unknown contributors can be obtained by specifying their indices via the optional argument `U`. If desired, customised summaries can be computed directly from the networks, for instance by using the tools provided in the RHugin package.

The best configurations considering the entire range of alleles is obtained by specifying `type = "all"`. Firstly we condition on the observed peak heights, and then we request all genotypes with posterior probability above `pmin = 0.008`.

```
> setPeakInfo(mix15p4, ml15p4$mle)
> mpall <- map.genotypes(mix15p4, pmin = 0.008,
                         markers = "TH01", type = "all")
> mpall

$TH01
   n_1_1 n_1_2 n_1_3 n_1_4 n_1_5 n_1_6 n_1_7 n_1_8        Prob
1      0     1     0     0     0     0     0     1 0.162580086
2      0     0     1     0     0     0     0     1 0.154020884
3      0     0     0     0     0     0     0     2 0.138320077
4      0     1     1     0     0     0     0     0 0.089487581
5      0     0     0     0     1     0     0     1 0.075713215
6      0     0     0     1     0     0     0     1 0.067593129
```

```
7      0     2     0     0     0     0     0     0 0.046503149
8      0     1     0     0     1     0     0     0 0.044049830
9      0     0     2     0     0     0     0     0 0.042332290
10     0     0     1     0     1     0     0     0 0.041730780
11     0     1     0     1     0     0     0     0 0.039325576
12     0     0     1     1     0     0     0     0 0.037247596
13     0     0     0     1     1     0     0     0 0.018309818
14     0     0     0     0     2     0     0     0 0.010107153
15     0     0     0     2     0     0     0     0 0.008152291

attr(,"U")
[1] 1
attr(,"pmin")
[1] 0.008
attr(,"ptotal")
     TH01
0.9754735
attr(,"class")
[1] "map.genotypes" "list"
```

A summary method converts the predicted allele counts to a list of configurations of genotypes to faciliate further interpretation.

```
> summary(mpall)

TH01:
     U1.1   U1.2   Prob
1    6      9.3    0.162580
2    7      9.3    0.154021
3    9.3    9.3    0.138320
4    6      7      0.089488
5    9      9.3    0.075713
6    8      9.3    0.067593
7    6      6      0.046503
8    6      9      0.044050
9    7      7      0.042332
10   7      9      0.041731
11   6      8      0.039326
12   7      8      0.037248
13   8      9      0.018310
14   9      9      0.010107
15   8      8      0.008152

Total probability: 0.9755
```

The default (`type = "seen"`) is to only predict alleles that are seen in at least one EPG in the model, thereby avoiding the dispersed distribution that is a natural consequence of the uncertainty about where a dropped-out allele could be.

```
> setPeakInfo(mix15p4, ml15p4$mle)
> mp <- map.genotypes(mix15p4, pmin = 0.008,
                      markers = "TH01")
> summary(mp)

TH01:
     U1.1   U1.2   Prob
1    9.3     NA    0.247508
2    7      9.3    0.154021
3    9.3    9.3    0.138320
4    7       NA    0.136297
5     NA     NA    0.108665
6    8      9.3    0.067593
7    8       NA    0.059864
8    7      7      0.042332
9    7      8      0.037248
10   8      8      0.008152

Total probability: 1
```

Here, `NA` indicates a dropped-out allele, thus in the set of unobserved alleles $\{5, 6, 9, 10, 11\}$. Investigation of what the dropped-out allele might be can be done, for instance, by prediction of just the set of unseen alleles using `type = "unseen"`.

7.4 Simulation

Configurations of genotypes – or, more generally, configurations of any subset of nodes in the network – can be sampled by using Bayesian network techniques, for instance `simulate` in the RHugin package.

Peak heights can be sampled using `rPeakHeight`. The simulation works by first sampling a set of DNA profiles from the distribution represented by the networks and then sampling a set of peak heights conditionally on the set of DNA profiles. Note that, in the current implementation, the sampled genotypes are not stored. There are two choices of peak height distributions, from which to sample: `"conditional"` sampling conditionally on observed peaks at all other alleles, and `"joint"` sampling from the joint distribution of all alleles in the EPG – specifically then also marginally.

```
> rph <- rPeakHeight(mix15p4, ml15p4$mle,
                     markers = "TH01", nsim = 3)
> drop(rph$TH01)

          [,1]     [,2]      [,3]
[1,]   0.00000   0.0000   0.00000
[2,]  78.00944   0.0000   0.00000
[3,] 719.71654 874.3810 682.60793
[4,] 666.80617 784.0902 750.02010
[5,]  95.18605   0.0000  58.61788
[6,]   0.00000   0.0000   0.00000
[7,]   0.00000   0.0000   0.00000
[8,] 262.51093 306.3277 200.59182
```

7.5 Graphics

7.5.1 Electropherograms

The `plot` method for `DNAmixture` objects provides a plot of the peak heights in the form of stylized EPGs. Here, there is one plot per marker in a specified set of `markers` with the horizontal axis corresponding to repeat number for the alleles. The plots are highly customisable; for more details, see the help page for `plot.DNAmixture`. The following code produces Figure 7.1.

```
> par(mfrow = c(1,3))
> plot(mix15p4, markers = c("D19S433", "TH01", "FGA"))
```

For mixtures where a set of dyes is specified, it can be useful to use the dyes for a layout mimicking the original EPG (Figure 7.2). As an example we use the dyes for the SGM plus kit, but as MC15 does not include Amelogenin – and we do not want an empty plot for this marker – we exclude Amelogenin from the set of dyes.

```
> data(SGMplusDyes)
> dyes <- SGMplusDyes
> dyes
```

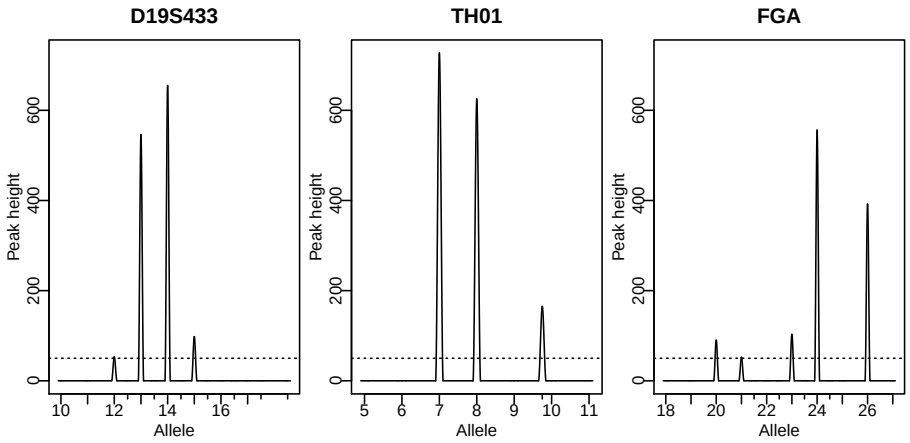


Figure 7.1: Electropherogram.

```
$blue
[1] "D3S1358" "VWA"     "D16S539" "D2S1338"

$green
[1] "AMEL"    "D8S1179" "D21S11"  "D18S51"

$yellow
[1] "D19S433" "TH01"    "FGA"

> dyes$green <- dyes$green[dyes$green != "AMEL"]
```

Now set the dyes for the mixture, and plot the EPG.

```
> dyes(mix15p4) <- list(dyes)
> plot(mix15p4, epg = TRUE,
       dyecol = list(c("blue", "green", "black")), font.main = 1)
```

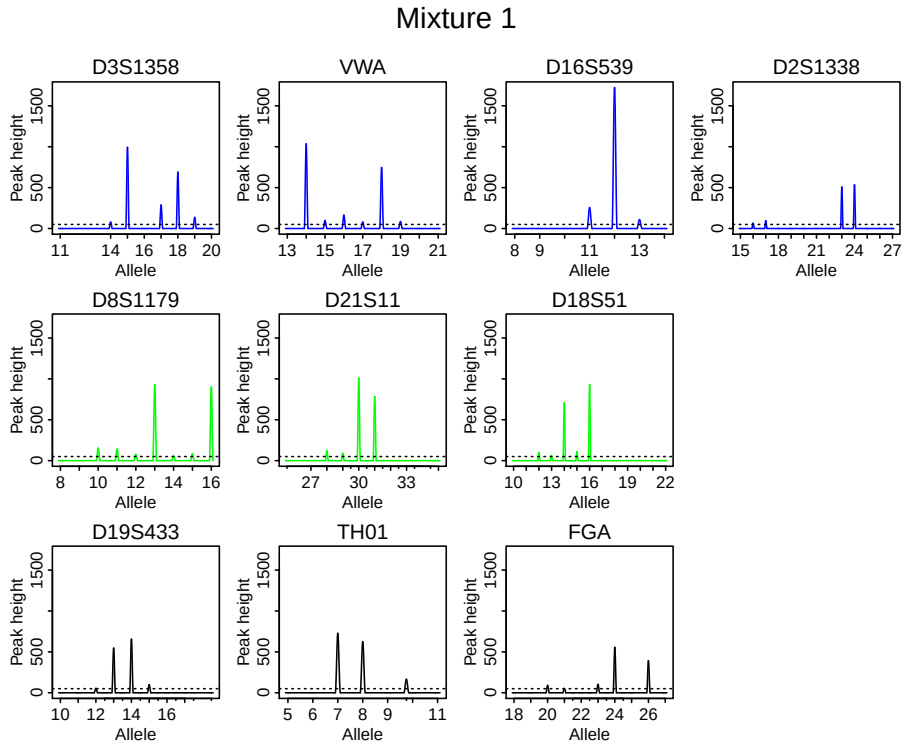


Figure 7.2: Electropherogram.

7.5.2 Diagnostic plots

Boxplots

A set of simulated peak heights can be visualised via the function `boxplot`. This plots the observed peak heights in the form of an EPG, with a boxplot for each of the peaks. As explained in Section 6.3, only the part of the boxplot that raises above the detection threshold is shown.

Below, we simulate from the distribution of the peak height at each allele conditionally on the observed peak heights at all other alleles, resulting in Figure 7.3.

```
> rph <- rPeakHeight(mix15p4, ml15p4$mle,
                     dist = "conditional",
                     markers = c("D19S433", "TH01", "FGA"),
                     nsim = 1000)
> par(mfrow = c(1,3))
> boxplot(mix15p4, rph, markers = c("D19S433", "TH01", "FGA"))
```

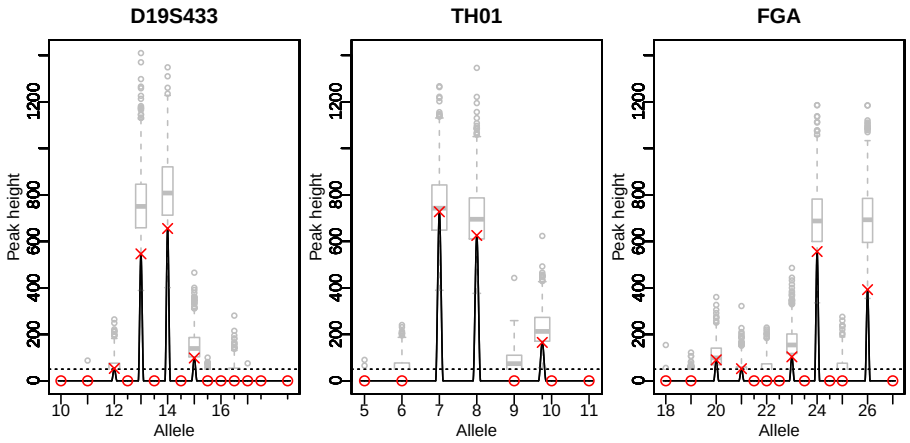


Figure 7.3: Simulated peak heights conditionally on the observed peak heights for all other alleles. The observed peak heights are indicated by red, with circles indicating a peak height of zero.

Quantile-quantile plots

The function `qqpeak` gives a quantile-quantile plot for assessing the peak height distributions under a given model. The possible peak height distributions are `joint`, `conditional`, and `prequential` corresponding to the probabilities obtainable by `predict`. The following code produces the quantile-quantile plots seen in Figure 7.4.

```
> par(mfrow = c(1,2), font.main=1)
> qqpeak(mix15p4, ml15p4$mle, dist = "conditional", main = "MLE")
> abline(0,1)
> qqpeak(mix15p4, psi, dist = "conditional", main="psi")
> abline(0,1)
```

Note that `qqpeak` invisibly returns a `data.frame`, which contains the probability transforms and other useful quantities.

Prequential monitors

The prequential score and other quantities needed for the prequential monitor can be computed through the function `prequential.score`. A plot method for prequential scores allows the visualisation of the score in comparison to approximative upper predictive limits (Figure 7.5).

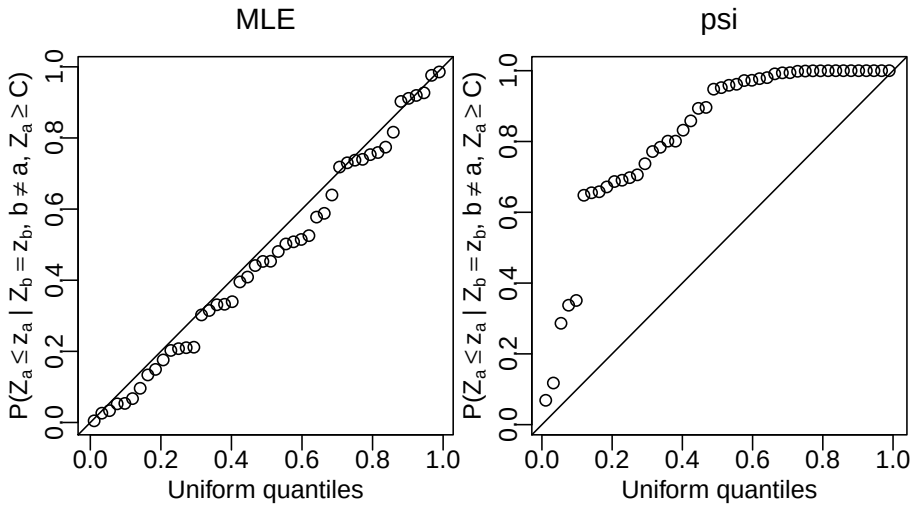


Figure 7.4: Quantile-quantile plots for the MLE (left) and for the arbitrary starting point `psi` used in the maximisation above (right).

By specifying a set of markers – here in the order of the dyes in the EPG – we can control the order in which they occur in the prequential monitor. Choosing only subsets of markers or DNA samples in the model is possible.

```
> preq <- prequential.score(mix15p4, ml15p4$mle, markers = unlist(dyes))
> plot(preq)
```

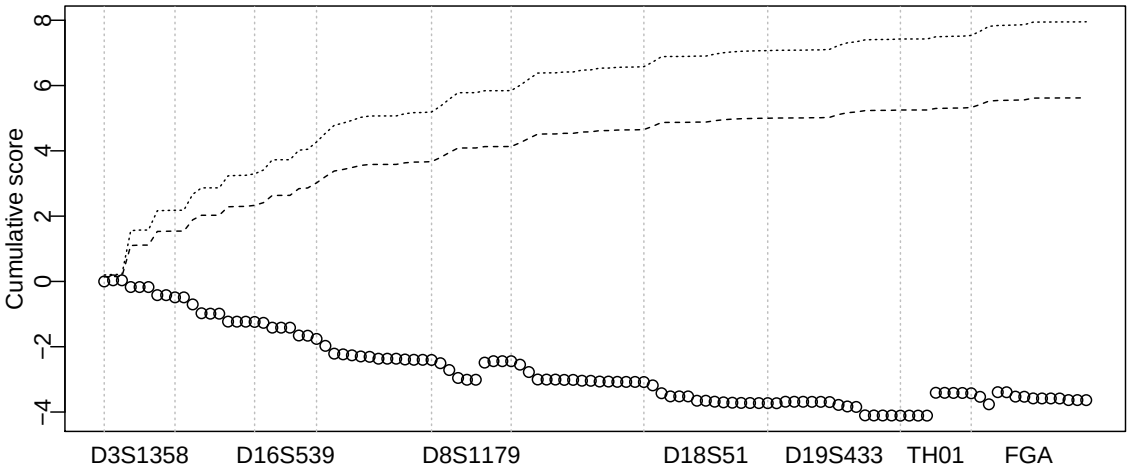


Figure 7.5: Prequential monitor.

The `prequential.score` object is a `data.frame` containing various quantities that are useful for making more customised plots. Particularly, it contains the logarithmic score `Y` with mean `EY` and variance `VY` as well as the cumulative `score` which equals `cumsum(Y-EY)`. The `prequential.score` also contains variables `marker`, `allele`, `height`, and `trace` (index of mixture included in the model) that can be used for identifying specific observations in the prequential monitor.

To investigate the upward jumps, in Figure 7.7 we have used red crosses to highlight all upward jumps and added a label with the allele name.

```
> par(mar=c(5.1, 2.1, 2.1, 0.1))
> ## Mark each upward jump by a red cross
> plot(preq,
       col = ifelse(preq$Y-preq$EY > 0, 2, 1),
       pch = ifelse(preq$Y-preq$EY > 0, 4, 1),
       las = 3 ## vertical axis labels
       )
> ## Labels for upward jumps
> text(preq$score, labels = preq$allele, pos = c(1,3),
       col = (preq$Y-preq$EY > 0), cex = 0.6)
```

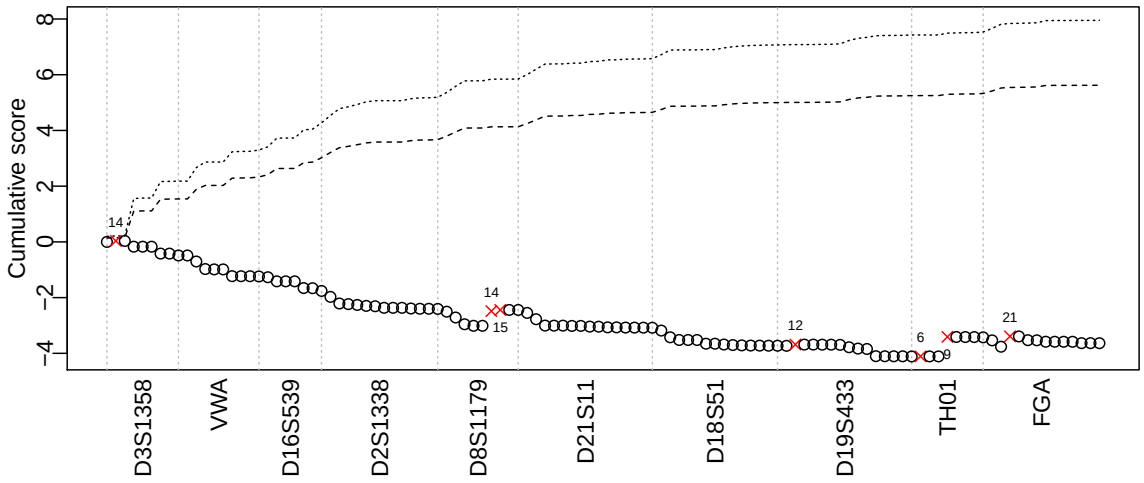


Figure 7.6: Prequential monitor.

The `prequential.score` object is a `data.frame` containing various quantities that are useful for making more customised plots. Particularly, it contains the logarithmic score `Y` with mean `EY` and variance `VY` as well as the cumulative

`score` which equals `cumsum(Y-EY)`. The `prequential.score` also contains variables `marker`, `allele`, `height`, and `trace` (index of mixture included in the model) that can be used for identifying specific observations in the prequential monitor.

To investigate the upward jumps, in Figure 7.7 we have used red crosses to highlight all upward jumps and added a label with the allele name.

```
> par(mar=c(5.1, 2.1, 2.1, 0.1))
> ## Mark each upward jump by a red cross
> plot(preq,
       col = ifelse(preq$Y-preq$EY > 0, 2, 1),
       pch = ifelse(preq$Y-preq$EY > 0, 4, 1),
       las = 3 ## vertical axis labels
       )
> ## Labels for upward jumps
> text(preq$score, labels = preq$allele, pos = c(1,3),
       col = (preq$Y-preq$EY > 0), cex = 0.6)
```

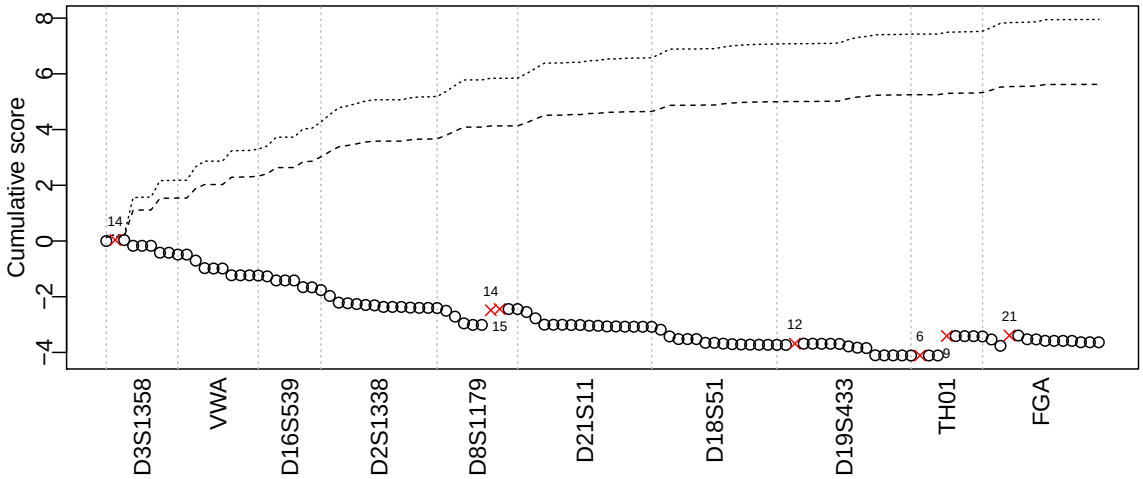


Figure 7.7: Prequential monitor.

Chapter 8

A case analysis: Four replicates of a low-template single-source DNA sample

The Netherlands Forensic Institute has kindly provided four DNA profiles, where the ground truth of the origin of the profiles is known: The profiles are four replicate analyses of a low-template single-source sample. In the following we shall ignore this knowledge and investigate what information the observed peak heights can provide about the profiles in terms of the number of contributors and their DNA profiles.

Observed peak heights

Peak heights are recorded for 15 STR loci plus Amelogenin and are displayed in Figure 8.1; the peak heights are found in Appendix A.2. The loci analysed correspond to those found in the NGM plus kit.

The specific kit used only affects the analysis below by determining the range of alleles for which peak heights can be observed. Other kit-specific values such as fragment lengths or parameter values have not been used.

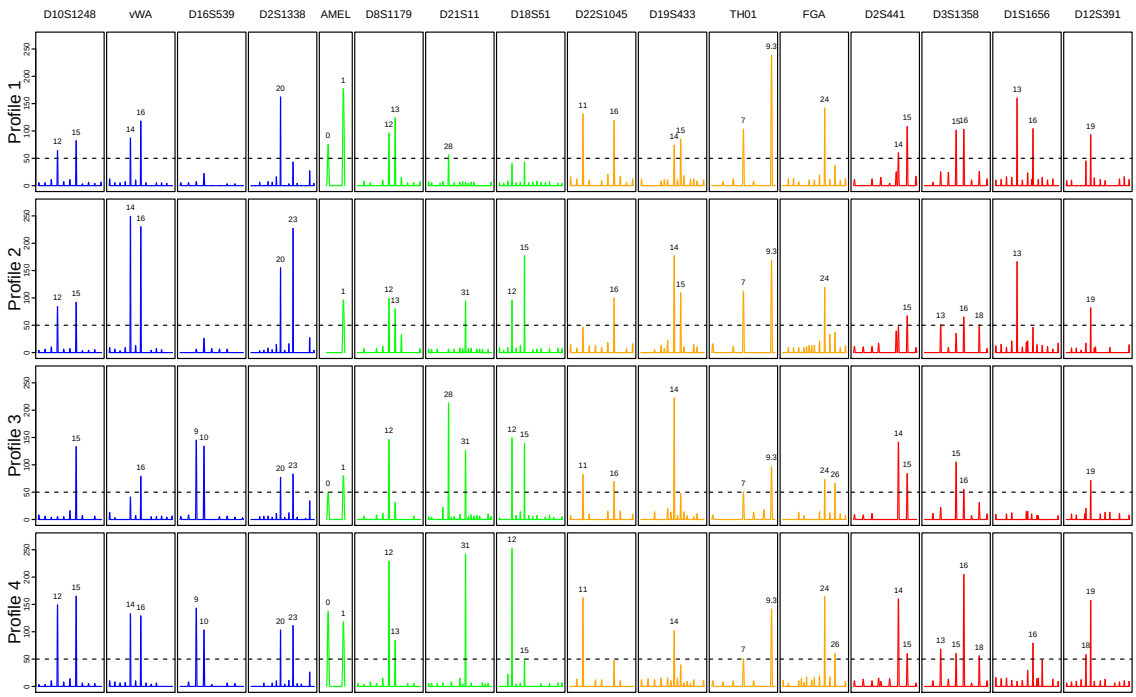


Figure 8.1: Peak heights for the four profiles. Peaks above the detection threshold (dashed lines) of 50 RFU are labelled by the repeat number of the corresponding allele.

The peak heights are in the range 2-252, consistent with the profiles being low template, and the ranges are similar across the four profiles. We use a detection threshold of 50 RFU, corresponding to the threshold used at the NFI, and hereby eliminate the many small peaks that we are worried are mostly noise. The cost is a loss of the information contained in any small allelic peaks and stutter peaks. For some of the profiles, markers D16S539 and D1S1656 do not exhibit any peaks above threshold.

We notice a significant overlap of observed alleles in the four profiles in Figure 8.1, indicating that there may be an overlap of contributors. Profiles 1 and 3 exhibit two peaks at all markers, and could a priori be explained by a single contributor. The peaks (13, 16, 18) and (13, 15, 16, 18) for marker D3S1358, profiles 2 and 4, can under our model only be explained by the presence of at least three distinct alleles (13, 16, 18) – allele 15 could be stutter – and thus at least two contributors.

There is no obvious sign of degradation in terms of a decline in peak heights with longer fragment lengths, and so it seems appropriate to continue the analysis without correcting for this.

Allele frequencies

We use the US-Caucasian allele frequencies from Budowle et al. (2011) also available in DNAmixtures, though note that an informed choice of reference population would be preferable.

Some of the alleles in the raw data for the four samples are not observed in our reference population, and furthermore some alleles are just outside the range of the NGM allelic ladder. We include all such alleles in the analysis, and correct for zero frequencies as follows. The published frequencies are based on 349 individuals. For each marker, if an allele has been seen in at least one EPG, we add one such allele to the database. After adding n alleles to the database for a particular marker, the allele frequencies are updated to reflect that the database is based on $2 \times 349 + n$ alleles at that marker. It should be noted that, since all of the added alleles exhibit peaks below the 50 RFU threshold, an alternative analysis could be carried out without adding any alleles to the database; we refrain from doing this.

8.1 Single profile analyses

In determining the origin of the four profiles, part of our task is to investigate whether they have any contributors in common. Firstly, let us consider each profile separately.

We start by explaining each profile as a mixture of two unknown contributors. The estimates are seen in Table 8.1 and are fairly similar across profiles. The parameter ρ is small, which indicates that the samples are low template, as does the the (generic) mean peak height of $\mu = \rho\eta \approx 100$ RFU. The profiles 1 and 3 are well

explained by a single contributor, whereas profiles 2 and 4 seem to have a second contributor accounting for 12% and 15% of the DNA, consistent with the presence of additional peaks at marker D3.

Table 8.1: Maximum-likelihood estimates for each profile separately. The combined likelihood for a model where the contributors are unrelated between profiles is -752.4.

Profile	ρ	η	ξ	ϕ_{U1}	ϕ_{U2}	$\log L$
1	3.591	25.760	0.031	1.000	0.000	-170.4
2	2.086	51.643	0.079	0.881	0.119	-185.4
3	4.142	21.601	0.008	1.000	0.000	-175.6
4	4.902	27.702	0.126	0.853	0.147	-221.1

The prequential monitors in Figure 8.2 are constructed so that within each marker the alleles are ordered by increasing length, and for each allele there are four points corresponding to the four EPGs. The prequential monitors stay below the upper prediction limits except in the plot for sample 2, where the model seems to exhibit some problems in predicting the present alleles. The quantile-quantile plots have few points as there are few observed alleles for each DNA sample, but Figure 8.2 gives no reason to believe that the profiles are inadequately explained by two or less contributors, and so we proceed to a joint model for the four profiles.

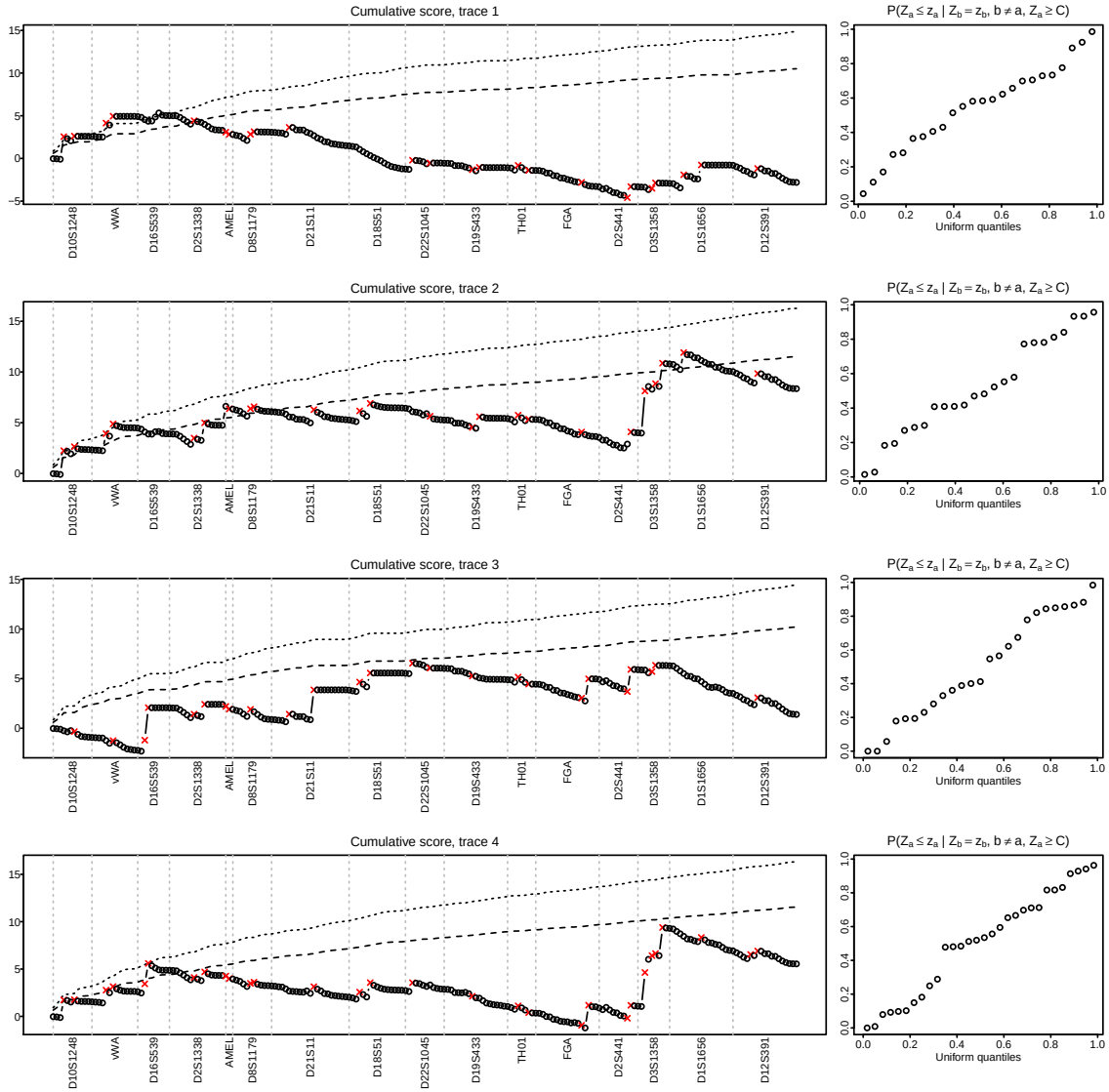


Figure 8.2: Diagnostic plots for the four models of Table 8.1 explaining each EPG separately by two unknown contributors. Red crosses indicate were above-threshold peaks has been observed.

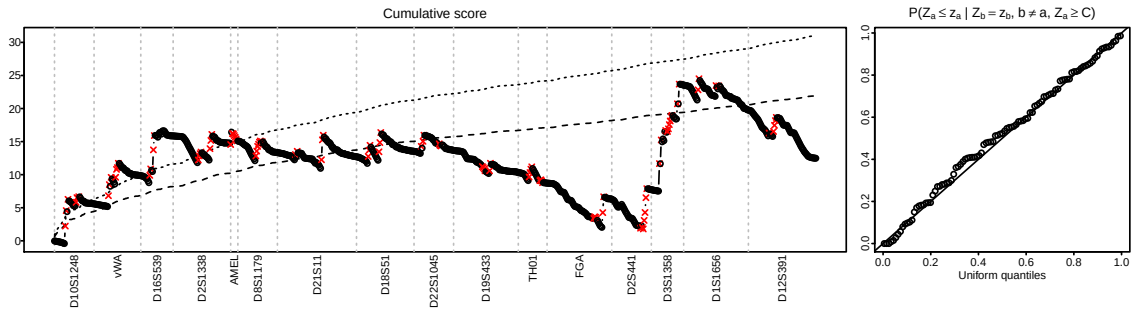
8.2 Joint analyses

The four separate analyses have pointed to a model with a joint set of at most 6 contributors being adequate. Simply adding the maximised likelihoods corresponds to assuming no overlap of contributors between mixtures; the latter is specified by a restriction that each contributor i has a contribution ϕ_{ei}, which is positive for one mixture e only.

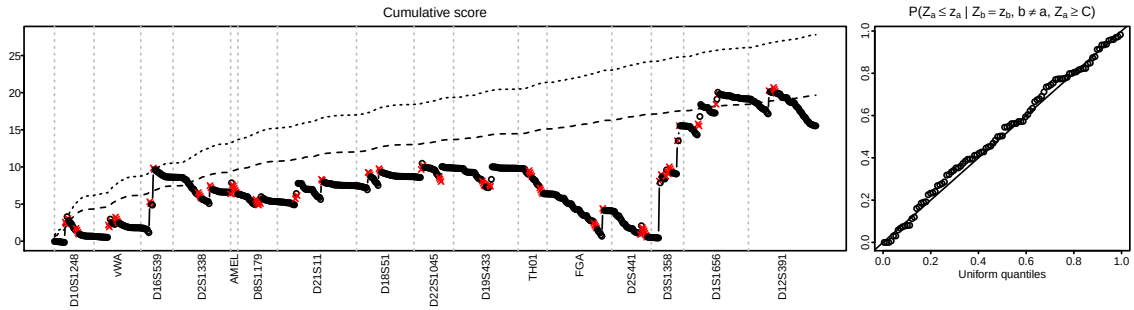
If the contributors between the profiles are indeed independent, so are the predicted probabilities for peak height events between the four profiles. Thus, we may combine the quantities used for the diagnostics in Figure 8.2 and assess the joint model. The resulting plots are seen in Figure 8.3a. There is clearly some trouble in predicting the presence of peaks; the observed peaks are independent between EPGs under this model, and therefore observing a peak in one EPG does not make it any more likely also to observe the peak in the next EPG – we see this effect for alleles that are observed in multiple EPGs by a series of successive jumps upwards.

Ideally we would now proceed to fit a model with a total set of 6 contributors and let the estimation point to the relevant configuration of the overlap of contributors between mixtures. Since six unknowns would require a significant computational effort, cf. Section 4.4, we fit the computationally less demanding model with a total of 4 unknown contributors, still allowing the possibility of only a partial overlap of contributors.

The diagnostic plots for the model with 4 shared contributors are seen in Figure 8.3b. As the contributors are now shared rather than independent, the model is no longer "surprised" by repeated occurrence of the alleles of the major contributor in the four EPGs. In Figure 8.3b there are two remarkable jumps in the prequential monitor at marker D3, corresponding to the peaks of the alleged minor contributor observed for the EPGs 2 and 4. The quantile-quantile plots wiggle slightly, but generally look fine.



(a) Two distinct contributors per crime-scene profile.



(b) Four shared contributors.

Figure 8.3: Diagnostic plots under two different joint models for the four crime-scene profiles. In the prequential monitors, red crosses mark peaks above threshold.

The estimates in Table 8.2 agree well with our previous findings: Profiles 1 and 3 share one contributor and the other three contributors are essentially not there. Profiles 2 and 4 have two common contributors in virtually common proportions, and all four profiles have the same major contributor. Thus, it would seem that a model with a total of two contributors is adequate; this reduction does, in fact, not change the value of the maximised likelihood. We notice that the estimates for ξ are now zero for some of the profiles.

Table 8.2: Estimates for four shared contributors. The maximised likelihood is $\log \hat{L} = -666.1$.

Profile	ρ	η	ξ	ϕ_{U_1}	ϕ_{U_2}	ϕ_{U_3}	ϕ_{U_4}
1	3.241	28.482	0.000	0.986	0.005	0.005	0.005
2	2.375	44.828	0.000	0.857	0.000	0.000	0.143
3	3.292	28.205	0.039	1.000	0.000	0.000	0.000
4	3.780	33.752	0.000	0.855	0.000	0.000	0.145

Predicted DNA profiles

In the joint model with just two contributors, the DNA profile of the main contributor U_1 to the mixtures is well determined, and the most likely profile – with a probability 1 at all but two markers – is indeed the profile of the actual donor to the four stains. There is some uncertainty in marker D19S433, where the genotype is (14, 15) with probability 99.7% and (14, 14) with probability 0.3%. The latter would evidently require U_2 to posses the allele 15 seen in the profiles. At marker D12S391 one allele is 19 and the other is – with probabilities (75.6%, 17.2%, 7.2%) – either 18, 19, or it has dropped out in all of the profiles.

Table 8.3: The posterior most likely profile for U_1.

D10S1248	vWA	D16S539	D2S1338
12, 15	14, 16	9, 10	20, 23
AMEL	D8S1179	D21S11	D18S51
XY	12, 13	28, 31	12, 15
D22S1045	D19S433	TH01	FGA
11, 16	14, 15	7, 9.3	24, 26
D2S441	D3S1358	D1S1656	D12S391
14, 15	15, 16	13, 16	18, 19

The distribution of the profile of the alleged minor contributor U_2 is quite diffuse. Further, the alleles of U_2 are almost all explained either by dropout or they are masked by the profile of U_1. One exception is at marker D12S391 where, if U_1 is homozygous with genotype (19,19) then the observed allele 18 is attributed to U_2 rather than to stutter. Another exception is at marker D19S433, where if U_1 is homozygous (14, 14) then U_2 must explain the observed peak at allele 15. Finally, at marker D3S1358 the minor contributor is needed to explain the presence of four peaks, $(13, 15, 16, 18)$, as these cannot be explained by a combination of a single contributor and stutter. Although the peak at allele 15 could be explained by stutter from allele 16, the model assigns with probability 1 the genotype $(15, 16)$ to U_1 and the genotype $(13, 18)$ to U_2.

8.3 Unexplained artefacts

Now, we know that there is indeed no second contributor; the peaks 13 and 18 for marker D3S1358 are explained by the laboratory at the NFI as interference with the internal standard, and the risk of such interference is one reason for the application of a threshold of 50 RFU.

There are various approaches for handling masking peaks in our model. One simple way, which is readily possible using DNAmixtures, is to exclude the entire marker for the two EPGs exhibiting these peaks or use a higher threshold for that marker. However, that would imply an unnecessary loss of information about other peaks. Another more satisfactory approach would be to exclude just the observed peak heights for the relevant alleles. A third possibility is to allow an allele-specific detection threshold and set this equal to the observed peak height, which would exploit the information that any masked peak is at most this size.

Ignoring marker D3S1358 for the profiles 2 and 4 and re-fitting the model with 2 shared contributors for the four mixtures points to the second contributor not being present; indeed the estimates of ϕ_{U_2} seen in Table 8.4 are in essence zero, and the maximised likelihood does not change at all by removing the second contributor.

Table 8.4: Estimates in the two-contributor model after excluding marker D3S1358

Profile	ρ	η	ξ	ϕ_{U_1}	ϕ_{U_2}
1	2.857	31.635	0.000	1.000	0.000
2	1.984	51.058	0.000	1.000	0.000
3	2.905	31.690	0.020	1.000	0.000
4	3.759	32.302	0.083	1.000	0.000

Our final model explains the samples as four replicate analyses of a single source profile, with no restriction of the parameters to be common across the four samples. The general model fit is fine as Figure 8.4 illustrates, although perhaps the quantile-quantile plot suggests some difficulty in predicting the heights of the observed peaks. One simple reason for this could be that the distribution of the unknown contributor does not describe the profile of the donor very well.

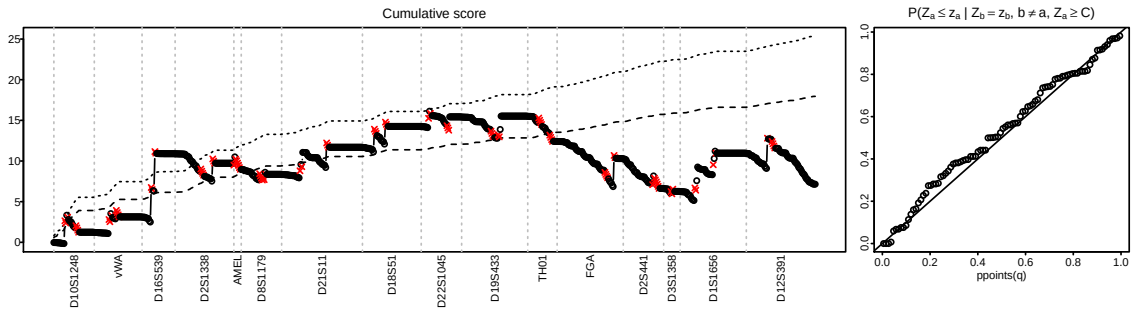


Figure 8.4: Diagnostic plots for the single contributor model.

The most likely profile still matches that of the actual donor. Based now only on profiles 1 and 3, the genotype for D3S1358 is again predicted to be $(15, 16)$. Some ambiguity remains in the allele designation at marker D12S391, where one allele is definitely 19, but the other allele is either allele 18 (51%), allele 19 (42%), or have dropped out completely (7%). Thus, the probability of U_1 being a homozygote has increased compared to the two-contributor model; in this event, the observed allele 18 would be attributed to stutter. As the peak height for allele 18 is as much as 37% the height of the peak at allele 19, it would not be classified as stutter in a typical analysis applying the common threshold of 15% for stutter peaks.

In principle, we can for any suggested profile of the donor investigate further whether it explains the data well, as this is simply a matter of assessing the fit of a model that includes the suggested profile as a known donor. We shall leave out such investigations.

The above analysis illustrates that our model can be useful for the analysis of both low-template DNA and stains from a single donor. As a point of reference, we note that the analysis has been performed using DNAmixtures and takes around 2.5 hours to run on a standard desktop. The majority of this time is spent on the maximisation in the joint model with four unknown contributors and depends heavily on the choice of starting point; in this case the starting point was chosen far

from the maximum point. The remaining maximisations and further analyses take around 20 minutes in total.

Chapter 9

Discussion

In this thesis, we have aimed to have a high level of modularity in our model and methodology. As a consequence, many aspects of the analysis can be changed without affecting the validity of the methodology.

In this chapter, we highlight three areas in which changes are likely to be desirable and give some suggestions as to how this may be achieved.

The first area concerns our choice of peak height distribution. In this connection we discuss briefly two other models, Cowell et al. (2011) and Puch-Solis et al. (2013), that use gamma distributions in the modelling of peak heights. We address the overall suitability of gamma distributions and outline a few suggestions regarding alternative choices of distributions.

The second area concerns the assumptions about the distribution of genotypes. We have assumed independence between the profiles of unknown contributors. It may be relevant to introduce dependence between the unknown contributors, for instance by modelling relatedness between contributors or by modelling the uncertainty of allele frequencies rather than treating them as fixed and known.

The third area that we touch upon concerns the way of handling unknown parameters. In practice, maximising the likelihood is a time-consuming part of a DNA mixture analysis and further investigation of alternative optimisation methods could

therefore be of great value. Finally, we discuss a few alternatives to maximum likelihood estimation.

9.1 Assumptions about peak height distributions for fixed genotypes

It should be noted that neither the computational methodology nor the further statistical inference are limited to our specific choice of peak height distributions.

One possible modification relates to the specific parameterisation of the gamma distributions. It is assumed that the peak height distributions can be adequately described by just four model parameters, and the question whether a more complex dependence of the parameters on alleles is necessary should be properly addressed.

More generally, the model does not rely on the peak heights being gamma distributed, and it may well be the case that there are better choices of families of distributions. Other sources have used Gaussian distributions (Cowell et al., 2007b; Tvedebrink et al., 2012b, 2010; Perlin et al., 2011) or log-normal distributions (Taylor et al., 2013) in modelling peak heights.

9.1.1 Marker- and allele-dependent parameters

We have considered all parameters to be constant across markers and alleles, but this might be too restrictive; Section 6.3.3 touched briefly on these issues. In its current form, DNAmixtures readily allows locus-specific parameters, although one should be aware that this naturally adds to the complexity of the maximisation of the likelihood.

Commonly, peak height distributions are adapted to capture degradation or preferential amplification leading to lower peak heights for longer fragment lengths. One such adjustment (Balding, 2013; Puch-Solis et al., 2013) is to scale the mean peak height by fragment lengths or a function thereof. Cowell et al. (2011) directly scale

the peak heights, which in their model – because they do not apply a detection threshold – is equivalent to a scaling of the mean peak height.

As discussed in Cowell et al. (2015), a simple way to model degradation would be based on the degradation model of Tvedebrink et al. (2012c). For this, we let ρ, which is thought of as proportional to the amount of DNA, be allele-specific and dependent on the fragment length λ_a as

$$\rho_a = \alpha\beta^{\lambda_a}.$$

As the fragment lengths are known, this adds just one extra parameter to the model.

Also, studies indicate that the stutter percentage increases with the length of the allele. Taylor et al. (2013) suggest that stutter increases with the *longest uninterrupted sequence* of repeated motifs, but highlight the difficulty in making such detailed models, as it requires knowledge about the specific composition of an allele rather than just the length.

9.1.2 Other models using the gamma distribution

The models of Cowell et al. (2011) and Puch-Solis et al. (2013) are both extensions of Cowell et al. (2007a) and thus similar to the model discussed in Chapter 2. Below we give a brief summary of their work and highlight in which ways it diverges from the work presented in this thesis.

Cowell et al. (2011)

The gamma model of Cowell et al. (2007a) was extended in Cowell et al. (2011) to allow for stutter, dropout, and silent alleles. The model can be used for both peak heights and peak areas; the model was described for a peak weight, which is the height or area scaled by repeat number as a simple correction for longer alleles having smaller peaks. The model requires user-specified parameter estimates and is thus not immediately operative. For their case analysis an informed albeit arbitrary

choice of parameters is used. The discretized vector ϕ of mixture proportions is given a uniform prior.

In Cowell et al. (2011), dropout is explained solely by pre-PCR dropout of alleles, and then the peak heights are gamma distributed conditionally on the surviving alleles. In contrast, our model explains dropout solely by a peak failing to raise above the detection threshold C; the variability in pre-PCR extraction phase – provided that no dropout occurs – may be seen as incorporated in the variability of the gamma distribution; this was discussed in Cowell (2009) and is further discussed in Section 9.1.3 below.

For each of a person's two alleles independently, all or no copies are put forward for PCR according to a dropout-probability $\delta_i = \exp(-\lambda\gamma\phi_i)$, where the probability of an allele dropping out decreases with its amount in the mixture. In their case analysis $\lambda\gamma \approx 4$. This model for pre-PCR dropout corresponds to using modified allele counts n^*_{ia} that are sampled from a binomial distribution with count n_{ia} and probability of selection $1 - \delta_i$. Mimicking our Bayesian network representation of allele counts (Figure 4.1), the modified allele counts can be represented by the network in Figure 9.1. Although the all-or-nothing dropout model of Cowell et al.

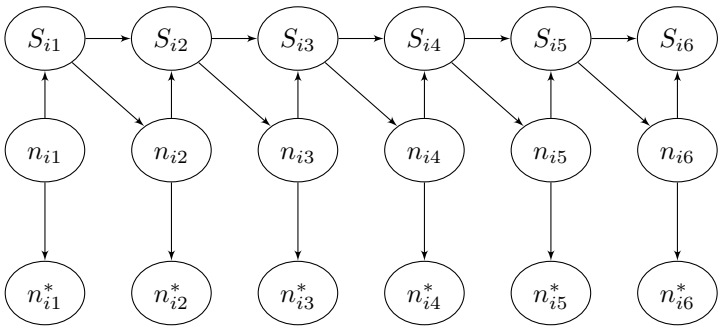


Figure 9.1: Modified genotype network, where only some alleles are sampled.

(2011) is somewhat crude, it offers a possible advantage in the ability to explain extreme dropout in cases where the tail probability of a gamma distribution would not be large enough.

A random stutter fraction s_a is incorporated for each allele to modify the gamma shape parameter much like our (fixed) ξ. The fractions s_a are assumed mutually independent and independent of the genotypes. For their case analysis, each s_a takes values $\{0, 0.05, 0.1\}$ with probabilities $(0.98, 0.01, 0.01)$; a coarse discretisation of s_a is necessary to render their computations feasible.

Conditionally on stutter fractions $\boldsymbol{s}$ and modified allele counts $\boldsymbol{n}^*$, the peak heights are mutually independent and gamma distributed as

$$H_a \,|\, \boldsymbol{s}, \boldsymbol{n}^* \sim \Gamma\left\{\rho\left((1-s_a)\sum_i \phi_i n^*_{ia} + s_{a+1}\sum_i \phi_i n^*_{i,a+1}\right), \eta\right\}.$$

To facilitate comparison to Model 2.2, we have here changed the notation of Cowell et al. (2011) to match that of Chapter 2. The original formulation uses θ for the mixture proportions; the amplification parameter ρ was instead expressed as a product of other parameters making the proportionality with the amount of DNA explicit; and η denoted the rate rather than the scale parameter. In their analysis of MC15 and MC18, they used $\rho \approx 32$, which we note is similar in magnitude to the maximum likelihood estimates obtained using our model, see e.g. Table 5.3.

Silent alleles are handled by adding an unobservable allele, corresponding to our approach in Section 4.5. For the case example an unobservable allele is added and given probability 0.005.

Puch-Solis et al. (2013)

The basic gamma model for peak heights was also adopted in Puch-Solis et al. (2013) and extended to allow for stutter and dropout in the context of at most two contributors.

A peak height distribution is obtained by decomposing a peak into independent gamma-distributed contributions. There are two such contributions per individual, one contribution coming from correctly amplified alleles and one contribution coming from stutter from the allele one repeat larger. Since a common scale parameter

is assumed for all contributions within a locus, the peak height is in turn gamma distributed. For all the gamma distributions, the shape parameter includes a scaling by the allele length in base-pairs.

The decomposition of peak heights is essentially equivalent to the one we make in Cowell et al. (2015), where different shape parameters for stutter contributions and allelic contributions are assumed and obtained by decomposing the effective number of alleles according to proportions ξ and $1 - \xi$.

Because a proxy for the unknown total amount of DNA in terms of the total peak height is introduced along with the model specification, the exact distributional assumptions become a bit unclear. For instance, the assumption of peak heights being gamma distributed conditionally on the observed total peak height seems contradictory to the assumption that the scaled peak heights are then Dirichlet distributed.

Their dropout model is similar to ours in that dropout is modelled by introducing a detection threshold for the absolute peak heights. As they subsequently model relative peak heights the censoring becomes more complicated. The benefit of considering relative peak heights is unclear; an advantage of modelling relative peak heights as introduced in Cowell et al. (2007a) is that the scale parameter in the gamma distribution becomes redundant, but once a detection threshold is introduced, this advantage disappears.

Dropin is, as in our case, attributed to the presence of additional unknown contributors.

In order for the model to be operative, they introduce an approximation by grouping all unobserved alleles at a locus into one compound allele. The mixture proportion is discretized and given a uniform prior. Other model parameters are estimated from experimental data to each locus separately.

9.1.3 Branching processes and PCR

Some investigation of the suitability of the gamma model for peak heights was carried out in Cowell (2009) as a simulation study based on the model of Gill et al. (2005).

We may think of the EPG as obtained in the following three stages: in the first stage, a subset of the original sample is extracted and put forward for PCR; in the second stage, the PCR process amplifies the selected molecules; and finally in the third stage, the EPG is produced from the fluorescence intensity of the molecules detected during electrophoresis.

The simulation model of Gill et al. (2005) models the number of molecules after PCR by explicitly modelling the two first of the three stages. It is unclear how to extend such models to models for the signal in the EPG, as it seems plausible that some additional variability is introduced at this stage. Cowell (2009) assumes that the conversion to peak heights can be seen as a mere scaling, meaning that a good approximation of the gamma distribution for modelling the number of molecules post-PCR translates directly into a gamma model for the peak heights.

Although a simplistic model, the branching process model for PCR provides a framework that can be used to motivate or interpret peak-height models and it constitutes a useful tool in identifying suitable families of distributions. Below we study the distribution of the number of molecules after amplification by modelling the PCR process as a branching process starting from either a fixed or a random number of template molecules.

Modelling PCR by a branching process

Starting from N_0 template molecules, let N_t denote the number of molecules after t cycles of PCR. We assume that in each cycle, a molecule produces a (non-negative) random number of copies independently of other molecules, and that after each cycle the population consists of both template molecules and their copies. Firstly, assume that amplification starts from just one molecule, i.e. $N_0 = 1$.

We restrict attention to super-critical branching processes, for which $\mathbb{E}(N_1 \mid N_0 = 1) > 1$ corresponding to a strictly positive expected number of copies per molecule per cycle, and follow (Harris, 1963, Chapter 8) on the theory on such processes.

We may consider the observed number of molecules after t cycles scaled by the expected number of molecules $\mathbb{E}\, N_t = (\mathbb{E}\, N_1)^t$. The resulting process $N_t/(\mathbb{E}\, N_1)^t$ converges in mean square and in probability to an absolutely continuous random variable W. Note that the convergence in mean square implies convergence of expectations and variances, i.e. that $\mathbb{E}\, N_t/(\mathbb{E}\, N_1)^t \to \mathbb{E}\, W$ and $\mathbb{V}\, N_t/(\mathbb{E}\, N_1)^t \to \mathbb{V}\, W$. Rather than directly studying the distribution of the number of molecules N_t after t cycles, we consider the distribution of the limiting variable W. The mean and variance for the normalised number of molecules are

$$\mathbb{E}\, W = 1, \qquad\qquad \mathbb{V}\, W = \frac{\mathbb{V}\, N_1}{\mathbb{E}\, N_1(\mathbb{E}\, N_1 - 1)}.$$

Further, both the moment generating function (m.g.f.) and the characteristic function for W satisfy the functional equation

$$\phi(s\, \mathbb{E}\, N_1) = f(\phi(s)). \tag{9.1}$$

Importantly, this functional equation uniquely determines the distribution of W (Harris, 1948).

Many authors have adopted and discussed branching processes for modelling the PCR process; a few examples are Sun (1995); Stolovitzky and Cecchi (1996); Gill et al. (2005); Cowell (2009); Lalam et al. (2005); and Jo et al. (2011). Commonly, the following branching process has been used.

One or no copies. Let us firstly consider a branching process where each molecule would give rise to a copy with probability p – the *PCR efficiency* – and to no copy

with probability $1-p$, i.e.

$$\mathrm{P}(N_1 = i \,|\, N_0 = 1) = \begin{cases} 1-p, & i = 1 \\ \\ p, & i = 2. \end{cases} \tag{9.2}$$

In this setting, we have $\mathbb{E}\, N_1 = 1+p$ and $\mathbb{V}\, N_1 = p(1-p)$. The probability generating function for N_1 is

$$f(s) = \sum_{r=0}^{\infty} \mathrm{P}(N_1 = r) s^r = (1-p)s + ps^2, \tag{9.3}$$

and so – according to (9.1) – the moment generating function (as well as the characteristic function) for W satisfies the functional equation

$$\phi(s) = p\phi\left(\frac{s}{1+p}\right)^2 + (1-p)\phi\left(\frac{s}{1+p}\right). \tag{9.4}$$

Assuming that W is gamma distributed with shape λ and scale β, these parameters would be

$$\beta = \mathbb{V}\, W = \frac{1-p}{1+p}, \qquad\qquad \lambda = \mathbb{E}\, W / \mathbb{V}\, W = \frac{1+p}{1-p}. \tag{9.5}$$

The moment generation function for this gamma distribution is

$$\phi(s) = \frac{1}{(1-\beta s)^{\lambda}} = \frac{1}{\left(1 - \frac{1-p}{1+p} s\right)^{\frac{1+p}{1-p}}}.$$

Unfortunately, this does not satisfy the functional equation (9.4) – a numerical counterexample is easily found.

Although the above standard branching-process model for PCR does not enable an interpretation in terms of a gamma distributed limit, we shall investigate below whether the gamma distribution constitutes a reasonable approximation.

Evidently, the similarity of the limiting distribution and the gamma distribution depends on the exact branching process used for modelling the PCR process – as

we shall now see, there exists a simple branching process, for which the limit W is indeed gamma distributed.

Geometric offspring distribution. Making the slightly unnatural assumption that a molecule can produce an unlimited number of molecules in a single cycle, we let the number of molecules arising from a single template molecule follow a geometric distribution,

$$\mathrm{P}(N_1 = i | N_0 = 1) = (1-p)p^{i-1}, \ i = 1, 2, 3, \ldots$$

Under this model, there is a probability $1-p$ of a molecule not being copied ($N_1 = 1$), and a probability p of one or more copies ($N_1 \geq 2$); the probability of just one copy is $p(1-p)$ and probability p^2 of a molecule producing two or more copies in a single cycle. The moments for N_1 are $\mathbb{E}\, N_1 = \frac{1}{1-p}$ and $\mathbb{V}\, N_1 = \frac{p}{(1-p)^2}$ and the probability generating function for N_1 is

$$f(s) = \frac{s(1-p)}{1-sp}.$$

To respect that $\mathbb{E}\, W = \mathbb{V}\, W = 1$, if W is gamma distributed the shape and scale are both equal to 1. The moment generating function $\phi(s) = 1/(1-s)$ for this gamma distribution is seen to satisfy the functional equation (9.1):

$$\begin{aligned} f(\phi(s)) &= \frac{(1-p)\phi(s)}{1-p\phi(s)} = \frac{(1-p)\frac{1}{1-s}}{1-p\frac{1}{1-s}} = \frac{1-p}{1-s-p} = \frac{1}{1-\frac{s}{1-p}} \\ &= \phi\left(\frac{s}{1-p}\right) = \phi(s\,\mathbb{E}\, N_1). \end{aligned}$$

Thus, we conclude that the geometric offspring distribution leads to a gamma distributed limit.

In the remainder of this discussion, we adopt the standard one-or-no-copy model for the PCR process discussed above, where a molecule produces one copy with a probability p and no copies with probability $1-p$.

Identifying suitable families of distributions

In Gill et al. (2005), the probability p of a molecule producing a copy was estimated to 0.82 with a standard error of 0.12, indicating a relevant range of $p \in (0.58; 1)$. The extreme cases $p \in \{0, 1\}$ correspond to the deterministic scenarios where either no copies are produced or the PCR process is perfect.

Figure 9.2 shows the estimated densities for the limit W based on simulation of 28 PCR cycles and probabilities $p = 0.6, 0.82, 0.9$ of a molecule duplicating during a cycle. We see that for low PCR efficiencies the distribution of W is positively skewed

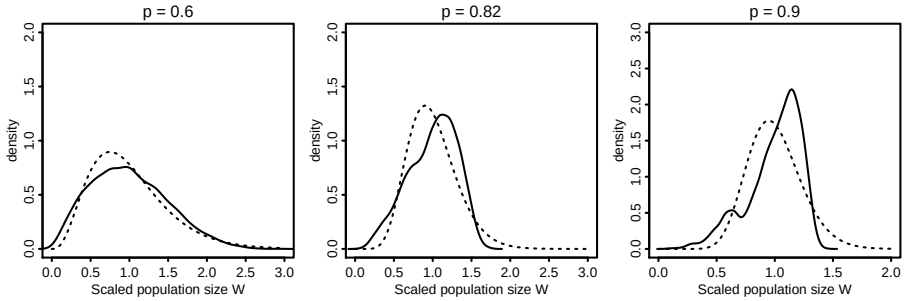


Figure 9.2: Distribution of the limit W (solid). The dashed curves correspond to the moment-matching gamma distributions (9.5).

and that as the efficiency increases the distribution becomes negatively skewed – a contrast to the positively skewed gamma distribution. The multimodal behaviour of the left tail for high PCR efficiency has to do with how many of the molecules fail to amplify during the early stages of the PCR-process and is discussed to some extent in Stolovitzky and Cecchi (1996) and Jo et al. (2011).

In Figure 9.3 we compare the limiting distribution to a set of standard families of distributions by comparing their skewness and (excess) kurtosis as a function of the PCR efficiency. In all cases, the distributions are chosen to match the mean and variance of W. Using K_i to denote the i'th cumulant, we define skewness as $K_3/K_2^{3/2}$ and kurtosis as K_4/K_2^2.

Note that the skewness and kurtosis for the distribution of W can easily be computed using the functional equation (9.1) for the moment generating function and that cumulants can be found from the derivatives of its logarithm.

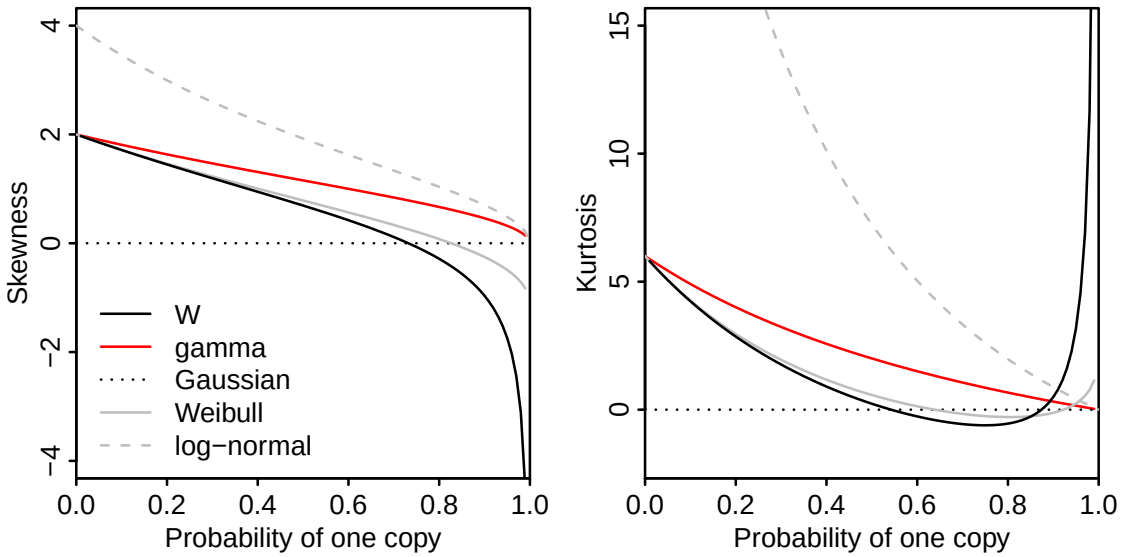


Figure 9.3: Skewness and kurtosis for the branching limit W (black) as well as a for a set of standard distributions.

For very high efficiency, the distribution of W becomes highly negatively skewed and with a large positive kurtosis. For high efficiencies p, the Gaussian distribution, with a skewness and kurtosis of zero, could potentially provide a better fit than the gamma distribution. For reference, we also compare the limiting distribution to the log-normal distribution, as this is used by Taylor et al. (2013); judging from Figure 9.3 it would not offer an improvement over the gamma distribution.

In contrast, the Weibull distribution has the property that it is negatively skewed for high efficiencies and Figure 9.3 suggests that the Weibull distribution could provide a good approximation; Jo et al. (2011) concluded that the approximation is good when the efficiency is low, i.e. $p \leq 0.5$, conforming well with Figure 9.3.

The gamma and Weibull distributions are both special cases of the generalised gamma distribution, which has an additional parameter α to the shape s and the scale η. The gamma distribution corresponds to $\alpha = 1$, and the Weibull distribution is the case where $\alpha = s$. An interesting study would be to substitute the gamma

distribution for peak heights with the generalised gamma, and investigate whether $\alpha = 1$ is suitable.

Multiple input molecules. In a more realistic setting where $N_0 \geq 1$ molecules are amplified, we assume that the PCR process can be seen as N_0 identically distributed branching processes evolving independently. This gives a total number $N_t = \sum_{i=1}^{N_0} N_{it}$ of molecules after t cycles, where each N_{it} is the result of a PCR process started from one molecule and N_0 is the number of molecules forwarded for PCR.

For each of the processes, we can consider the limit $W_i = \lim_{t\to\infty} N_{it}/\mathbb{E}\, N_{it}$ of scaled populations as before, as well as their sum $W = \sum_{i=1}^{N_0} W_i$, which – as a sum of (independent) limiting variables – is itself the limit of a scaling of the total number of molecules, $N_t/(\mathbb{E}\, N_{11})^t$. Note that the scaling is different to scaling with the expected *total* number of molecules after t cycles.

The skewness of the limiting distribution for N_0 molecules is $1/\sqrt{N_0}$ of the skewness when there is only one initial molecule and similarly the kurtosis is scaled by a factor $1/N_0$; effectively resulting in a vertical compression of Figure 9.3. This means also that the shape of the distribution will look more and more like the Gaussian. Figure 9.4 shows that the suitability of the gamma distribution does indeed seem to improve as the initial number of molecules to amplify increases.

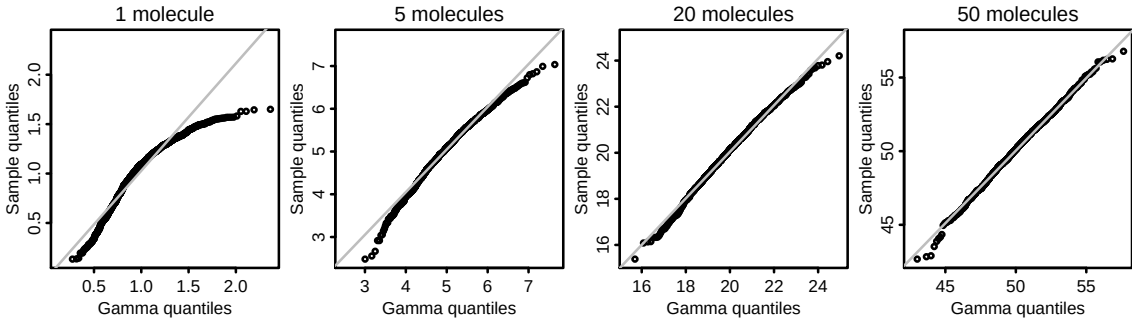


Figure 9.4: The scaled population size after 28 cycles using $p = 0.82$ and an initial input of $N_0 = 1, 5, 20, 50$ molecules. Lines through the first and third quartiles are in grey.

A random number of input molecules

We now turn to models where the initial number N_0 of molecules put forward for PCR is random, mimicking the randomness arising from the extraction procedure and thereby modelling explicitly the two first of the three stages of obtaining the EPG. When N_0 is random, the moments for W are

$$\mathbb{E}\, W = \mathbb{E}\, N_0, \qquad\qquad \mathbb{V}\, W = \mathbb{E}\, N_0 \left(\frac{1-p}{1+p} + \frac{\mathbb{V}\, N_0}{\mathbb{E}\, N_0} \right),$$

implying that if N_0 is proportional to the number of molecules in the sample, then so is the expected number of molecules after PCR. The moment matching gamma distribution is

$$\Gamma\left(\frac{\mathbb{E}\, N_0}{\frac{1-p}{1+p} + \frac{\mathbb{V}\, N_0}{\mathbb{E}\, N_0}}, \frac{1-p}{1+p} + \frac{\mathbb{V}\, N_0}{\mathbb{E}\, N_0} \right).$$

We notice here that if $\mathbb{V}\, N_0 / \mathbb{E}\, N_0$ does not depend on the number of molecules in the sample, nor does the scale parameter.

Gill et al. (2005) and Cowell (2009) both use binomial sampling to capture the pre-PCR extraction. To exclude the possibility of dropout in the extraction process – i.e. eliminate the possibility of $N_0 = 0$ – the simulation studies of Cowell (2009) further condition on a positive number of molecules being extracted.

A natural choice, also used by Tvedebrink et al. (2012a), is to model the extraction of some volume of the sample by a Poisson distribution. We let $\mathbb{E}\, N_0 = \lambda$, where λ is proportional to the number of molecules in the full sample. We simply shift the Poisson distribution rather than conditioning on $N_0 > 0$, thus letting $N_0 - 1$ be Poisson distributed with mean $\lambda - 1$. This implies that $\mathbb{E}\, N_0 = \lambda$ and $\mathbb{V}\, N_0 = \lambda - 1$. When $\lambda = 1$, we get a input of one molecule and thus a behaviour of the branching process limit as in Figure 9.3.

Figure 9.5 shows how the additional variability introduced in the number of input molecules affects the skewness and kurtosis of the branching process limiting distribution. The overall behaviour is more similar to the gamma distribution than

in Figure 9.3. In particular, unless λ is very close to 1, the limit distribution is positively skewed regardless of the PCR efficiency p.

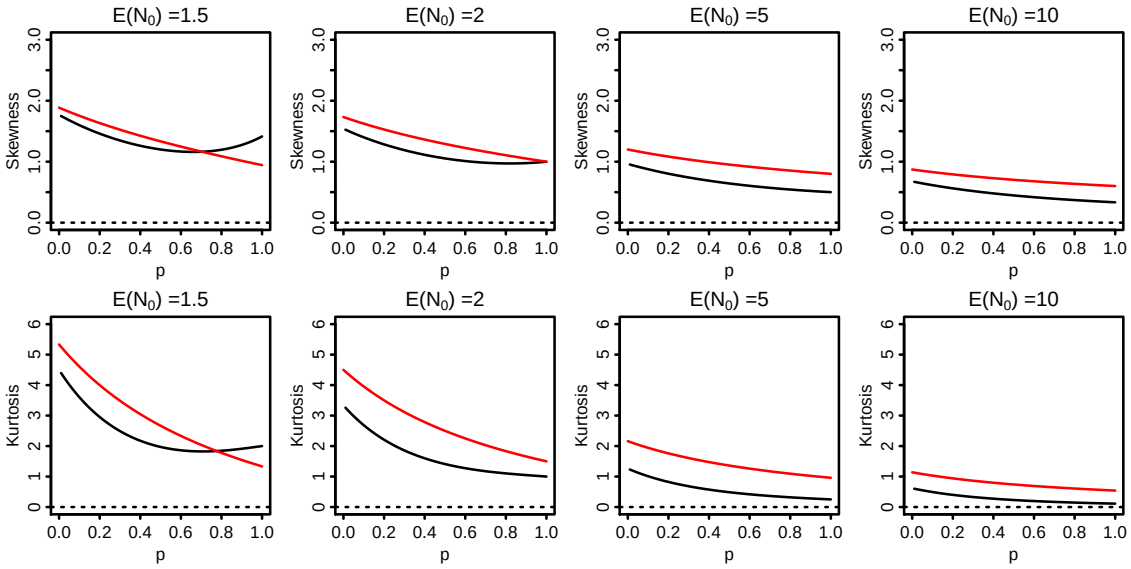


Figure 9.5: Skewness and kurtosis for the branching process limit W (black) and the moment matching gamma distribution (red) when modelling $N_0 - 1$ by a Poisson distribution.

9.2 Assumptions about genotype distributions

We have so far only discussed the scenario where the two alleles at a locus are sampled independently, and where all unknown contributors are assumed unrelated and unrelated to any known contributors. In practical applications, it is likely that the introduction of either close or distant relatedness is desirable.

In Proposition 2 we presented the conditional distribution (4.3) of n_{ia} given the partial allele count $S_{i,a-1}$. This distribution is based on a fixed set of allele frequencies and a total number of 2 alleles to allocate for the diploid case, i.e. $\sum_a n_{i,a} = 2$. Thus it holds that

$$n_{ia} \mid (S_{i,a-1}, \textstyle\sum_a n_{i,a}, \boldsymbol{q}) \sim \text{bin}\left(\sum_a n_{i,a} - S_{i,a-1}, q_a / \sum_{b \geq a} q_b\right).$$

This distribution depends only on the allele frequencies through q_a and $\sum_{b \geq a} q_b$. Knowing that allele frequencies sum to 1, we may also phrase the latter dependence

in terms of cumulative frequencies

$$S_a^q = \sum_{b<a} q_b = 1 - \sum_{b \geq a} q_b.$$

This leads to the idea of representing the distribution of allele frequencies themselves by a Markov structure similar to that of a genotype (Figure 4.1). When this is possible, then an unknown contributor depends on this subnetwork as shown in Figure 9.6. Note that this joint network also exhibits an overall Markov structure.

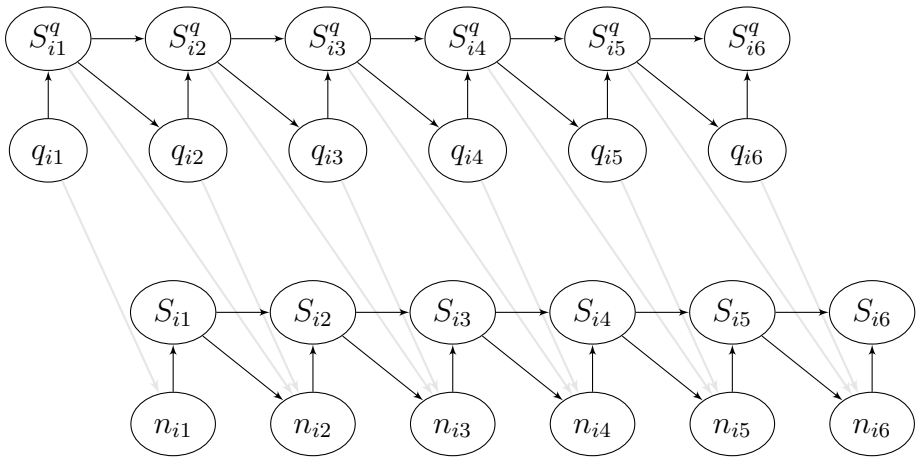


Figure 9.6: Network under a Markov representation of the distribution of allele frequencies.

Close relatedness

Considering a diploid marker, a child inherits two alleles independently, one from each of its two parents, and each allele is selected with equal probability among the two alleles of the parent. Conditionally on the allele counts $n_{i1}, \ldots, n_{iA}$ of a parent, the allele inherited by the child – as represented by a vector of allele counts – follows a multinomial distribution with frequencies $n_{i1}/2, \ldots, n_{iA}/2$. Importantly, as the allele counts exhibit a Markov structure, so do these (random) allele frequencies generated by the allele counts for the parent.

Using this parent-child model, we can then extend to sibling relationships by including also the shared parents and using that the genotypes of siblings are con-

ditionally independent given the genotypes of their shared parents.

More generally, it seems possible to introduce various forms of dependence between the alleles of unknown contributors whilst maintaining the overall Markov structure. This would enable an analysis along the lines of Green and Mortera (2009).

9.3 Handling unknown parameters

The estimation of unknown model parameters is an integral part of the DNA mixture analysis. Below, we discuss two aspects of estimation. The first addresses alternative methods for maximising the likelihood. The other aspect concerns the application of other estimation methods than the method of maximum likelihood.

9.3.1 Alternative optimisation methods

As an alternative to `solnp`, we can use the EM algorithm to perform the maximisation, in which case we will need to repeatedly maximise

$$Q(\psi \mid \psi') = \mathbb{E}\{\log f_\psi(\boldsymbol{Z}, \boldsymbol{n}) \mid \boldsymbol{Z}, \psi'\},$$

where the expectation is taken with respect to the conditional distribution of genotypes given observed peak heights and a current set of parameters ψ'. As $\log f_\psi(\boldsymbol{Z}, \boldsymbol{n})$ can be written as $\log f_\psi(\boldsymbol{Z} \mid \boldsymbol{n}) + \log p(\boldsymbol{n})$, and the distribution of $\boldsymbol{n}$ does not depend on the parameter ψ, we can maximise instead

$$\begin{aligned}
\tilde{Q}(\psi \mid \psi') &= \mathbb{E}\{\log f_\psi(\boldsymbol{Z} \mid \boldsymbol{n}) \mid \boldsymbol{Z}, \psi'\} \\
&= \mathbb{E}\Big\{ \sum_{m,a} \log f_\psi(Z_a^m, \boldsymbol{n}_a^m \mid \boldsymbol{n}_{a+1}^m) \,\Big|\, \boldsymbol{Z}, \psi' \Big\} \\
&= \sum_{m,a} \mathbb{E}\Big\{ \log f_\psi(Z_a^m \mid \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m) \,\Big|\, \boldsymbol{Z}, \psi' \Big\}.
\end{aligned}$$

This quantity may easily be computed using the network representing the distribution of genotypes conditionally on observed peak heights and the current parameter value ϕ', as obtained through conditioning on the auxiliary variables O_a. However, the maximisation of $\tilde{Q}(\psi \mid \psi')$ needs to be done by numerical methods. Thus, using the EM algorithm for maximisation only seems beneficial if the function Q can be maximised significantly faster than a direct maximisation of the full likelihood function.

Gradient-based optimisation methods require the partial derivatives to be either specified or computed numerically. For the likelihood function using gamma-distributed peak heights, exact derivatives by auxiliary variables requires differentiation of the incomplete gamma function in both shape and scale parameters; if an implementation of this is available, the gradient may be computed directly as

$$\begin{aligned}\frac{\partial}{\partial\psi}\log L(\psi \mid \boldsymbol{Z}) &= \frac{\partial}{\partial\psi}\tilde{Q}(\psi \mid \psi')\bigg|_{\psi'=\psi} \\ &= \sum_{m,a}\mathbb{E}\left\{\frac{\partial}{\partial\psi}\log f_\psi(Z_a^m \mid \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m)\,\middle|\,\boldsymbol{Z}, \psi\right\}\end{aligned}$$

where again the expectation is taken with respect to the conditional distribution of genotypes given observed peak heights and parameters ψ. The result is well known from the theory of the EM algorithm, see e.g. (McLachlan and Krishnan, 2008, p. 80). Note that many gradient-based methods will require a reparametrisation of the likelihood depending on the specified constraints on the parameters.

9.3.2 Alternative estimation methods

As the model complexity grows, there may not be enough information about the parameters in a single trace for estimation by maximum likelihood.

For the purpose of computing the WoE, it does not matter whether the parameters are well determined; a flat likelihood has a strong impact on the precision of the

parameter estimates, but not on the value of the maximised likelihood. However, the prediction of genotypes could well be sensitive to the parameters used.

For parameters relating to the PCR process and not to the specific mixture, prior information can be obtained from experimental data along the lines of other available models: Puch-Solis et al. (2013) and Mitchell et al. (2012) use maximum-likelihood estimates obtained from experimental data; Balding (2013) uses experimental data for the penalties in penalised maximum likelihood, so that parameters are still adapted to the case at hand; finally, Perlin et al. (2011) and Taylor et al. (2013) take a fully Bayesian approach and integrate out all unknown parameters.

Graversen and Lauritzen (2013) explored estimation of parameters in the basic gamma model of Cowell et al. (2007a) and discussed in this setting maximum likelihood, penalised maximum likelihood using prior information, as well as a fully Bayesian analysis.

For the fully Bayesian analysis of Graversen and Lauritzen (2013), a Gibbs sampler was used to alternate between sampling model parameters and genotypes, but as the genotypes and proportions of DNA from contributors are highly correlated, such a scheme exhibits some difficulty in properly exploring the state space.

A possible way forward is as follows. Denote by $\mathbf{g}$ the DNA profiles of the contributors to the mixture, denote by ψ the model parameters, and denote by E the available evidence. The evaluation of the likelihood $\Pr(E \mid \psi)$ is tractable, and so we have available the posterior ratio $\Pr(\psi \mid E)/\Pr(\psi' \mid E)$ which can be can be used in a Metropolis-Hastings scheme for sampling from the posterior distribution $\Pr(\psi \mid E)$. If desired, sampling from the joint model

$$\Pr(\mathbf{g}, \psi \mid E) = \Pr(\psi \mid E) \Pr(\mathbf{g} \mid \psi, E)$$

of parameters and DNA profiles $\mathbf{g}$ can readily be done using Bayesian network techniques to further sample from $\Pr(\mathbf{g} \mid E, \psi)$ as represented by the network of Graversen and Lauritzen (2014).

9.4 Final remarks

In this thesis, we have proposed a model for DNA mixtures and, in this framework, we have developed statistical and computational methodology for statistical analysis of DNA mixtures.

We have emphasised the importance of checking the adequacy of any model to the case data at hand, and the methodology and tools introduced for this can also assist in a systematic assessment of the general performance of the model. In the future, we hope to assess the performance at a larger scale using experimental data.

Although we have discussed the methodology in the light of the proposed model, we note that the considerations – particularly the diagnostic methods – are quite general.

A fundamental principle in our approach to DNA mixture analysis is that the model is developed separately from the further statistical and computational methodology. As a consequence, reasoning is ensured to be both consistent and transparent.

In practice, when using the implementation in DNAmixtures, it is ensured that we are working entirely within the adopted model. In particular, no further approximations are introduced; all computations are exact, apart from any imprecision arising from the use of numerical methods for maximisation and differentiation.

By providing the full implementation as an R package, we have attempted to demonstrate that it is indeed computationally feasible to use the model and methodology in practice. We hope that it enables and encourages a wider audience to challenge the suitability of the model and the general usefulness of the statistical methodology presented here.

Appendix A

Data

A.1 MC15 and MC18

Table A.1: Peak heights for DNA samples MC15 and MC18 together with genotypes of associated individuals.

(a) Blue dye

Allele	Peak height		Allele count		
	MC15	MC18	K_1	K_2	K_3
D3S1358					
14	79	50	0	0	0
15	993	715	1	0	0
16	0	67	0	0	0
17	286	479	0	2	1
18	689	638	1	0	0
19	135	136	0	0	1
VWA					
14	1036	876	1	0	0
15	98	249	0	0	1
16	163	274	0	2	0
17	79	97	0	0	0
18	746	967	1	0	0
19	85	251	0	0	1
D16S539					
11	256	534	0	1	1
12	1724	1786	2	1	0
13	109	265	0	0	1
D2S1338					
16	64	189	0	0	1
17	96	171	0	0	1
23	507	638	1	0	0
24	534	673	1	2	0

(b) Green dye

Allele	Peak height		Allele count		
	MC15	MC18	K_1	K_2	K_3
D8S1179					
10	152	241	0	0	1
11	140	192	0	0	1
12	76	127	0	0	0
13	929	1092	1	1	0
14	58	127	0	1	0
15	84	58	0	0	0
16	901	808	1	0	0
D21S11					
28	120	304	0	0	1
29	89	134	0	1	0
30	1010	1146	1	1	1
31	783	734	1	0	0
D18S51					
12	99	187	0	0	1
13	61	87	0	0	0
14	707	997	1	2	0
15	107	80	0	0	0
16	930	744	1	0	1

(c) Yellow dye

Allele	Peak height		Allele count		
	MC15	MC18	K_1	K_2	K_3
D19S433					
12	53	57	0	0	0
13	546	775	1	0	0
14	655	818	1	0	1
15	98	159	0	1	1
16.2	0	76	0	1	0
TH01					
7	727	670	1	0	0
8	625	636	1	0	0
9	0	99	0	2	0
9.3	165	348	0	0	2
FGA					
20	90	99	0	0	1
21	52	49	0	1	0
22	0	76	0	1	0
23	103	145	0	0	1
24	556	412	1	0	0
25	0	39	0	0	0
26	392	349	1	0	0

A.2 Low-template DNA from the NFI

Tables A.2-A.5 below show the peak heights for each of the four DNA profiles before a detection threshold is applied; $z_{e,a}$ denotes the peak height at allele a for the EPG $e = 1, \ldots, 4$.

Table A.2: Peak heights for markers in the blue dye lane.

(a) D10S1248

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
9	5	4	8	3
10	5	6	6	3
11	11	10	4	10
12	64	84	5	149
13	7	6	5	8
14	11	7	16	14
15	82	92	133	165
16	3	3	7	6
17	6	4	0	5
18	4	5	6	5
19	6	0	0	0

(b) vWA

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
10	12	9	13	10
11	5	6	3	8
12	5	3	0	6
13	7	9	0	7
14	87	249	41	133
15	10	13	7	10
16	118	230	79	129
17	5	0	0	6
18	0	4	5	4
19	5	7	5	6
20	5	5	6	0
21	4	0	4	0
22	0	0	6	0

(c) D16S539

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
7	5	0	5	0
8	5	0	8	8
9	7	6	145	143
10	22	26	134	103
11	0	7	0	3
12	0	6	5	0
13	3	6	6	6
14	3	0	4	5
15	0	0	3	0

(d) D2S1338

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
13	0	0	0	0
14	0	0	0	0
15	6	3	5	0
16	0	4	6	6
17	7	8	6	0
18	6	5	4	6
19	16	15	11	10
20	162	155	77	103
21	0	4	4	4
22	3	16	12	11
23	43	227	83	111
24	4	0	4	5
25	0	0	0	4
26	0	0	2	0
27	27	27	34	26
28	4	4	0	0

Table A.3: Peak heights for markers in the green dye lane

(a) Amelogenin

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
0	75	0	50	137
1	177	96	80	118

(b) D8S1179

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
7	0	0	0	5
8	8	7	6	3
9	5	0	0	8
10	0	7	8	5
11	10	11	11	15
12	96	99	146	229
13	124	80	31	84
14	15	32	0	0
15	5	0	0	5
16	5	0	6	5
17	7	7	0	0

(c) D21S11

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
24.2	7	6	5	5
25	5	6	4	5
26	0	6	0	5
26.2	4	0	0	0
27	7	0	22	6
28	56	6	213	0
28.2	0	0	4	8
29	5	6	5	0
29.2	0	0	0	0
29.3	0	0	0	0
30	6	8	9	15
30.2	7	7	0	4
31	6	94	126	242
31.2	4	7	5	0
32	6	8	8	6
32.2	6	0	6	0
33	0	6	7	0
33.2	0	6	6	0
34	0	6	0	7
34.2	0	0	0	4
35	0	5	10	5
35.2	6	0	5	0

(d) D18S51

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
9	5	9	3	0
10	4	4	0	0
11	8	9	0	22
12	40	96	149	252
13	5	8	7	4
14	6	13	14	6
15	43	177	139	50
16	5	0	7	0
17	6	5	5	0
18	8	6	7	0
19	6	7	0	5
20	5	0	4	0
21	7	7	8	5
22	0	0	4	0
23	4	8	0	6
24	4	7	5	6

Table A.4: Peak heights for markers in the yellow dye lane.

(a) D22S1045

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
9	16	15	7	0
10	12	8	0	13
11	131	46	82	162
12	10	12	10	0
13	0	13	0	12
14	9	9	0	12
15	20	18	15	0
16	119	100	69	47
17	17	0	15	10
18	6	7	0	0
19	12	16	9	0

(b) D19S433

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
9	12	0	0	12
10	0	0	0	14
11	0	5	14	12
12	8	13	0	15
12.1	0	0	0	0
12.2	11	7	0	0
13	10	22	20	15
13.2	0	0	13	10
14	74	177	222	102
14.2	11	0	7	15
15	85	109	47	39
15.2	18	10	14	6
16	0	0	7	9
16.2	11	0	0	5
17	12	14	5	12
17.2	8	10	10	0
18	0	0	0	0
18.2	10	0	0	11

(c) TH01

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
4	0	16	8	10
5	7	0	0	8
6	12	11	0	10
7	103	112	50	51
8	8	0	13	10
9	0	0	18	0
9.3	238	169	97	141
10	0	0	0	0

(d) FGA

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
16	0	0	0	11
17	12	9	0	5
18	13	9	0	0
19	7	9	13	10
19.2	0	0	0	14
20	0	9	7	7
20.2	0	11	0	17
21	10	13	0	0
21.2	0	12	0	11
22	10	13	0	16
22.2	0	0	0	0
23	19	21	14	19
23.2	0	0	0	0
24	142	119	73	164
25	11	33	12	17
26	37	37	66	60
27	9	9	10	0
28	13	12	7	9

Table A.5: Peak heights for markers in the red dye lane.

(a) D2S441

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
9	11	11	9	10
10	0	10	8	13
11	12	11	11	10
11.3	0	17	0	15
12	15	0	0	9
12.3	0	0	0	0
13	4	0	0	14
13.3	25	39	0	0
14	60	48	141	160
15	108	67	84	60
16	17	9	0	6

(b) D3S1358

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
11	0	0	0	0
12	6	0	11	10
13	25	51	22	68
14	24	9	0	13
15	101	35	105	60
16	103	65	55	205
17	10	0	7	6
18	26	51	31	56
19	12	7	10	10

(c) D1S1656

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
9	10	12	9	16
10	11	15	0	14
11	16	9	10	15
12	15	21	12	11
13	160	166	0	9
14	10	10	0	11
14.3	0	19	14	0
15	23	21	15	29
15.3	11	0	10	0
16	104	46	0	79
16.3	0	14	7	14
17	11	0	7	15
17.1	0	0	0	0
17.3	15	13	0	48
18	0	0	0	0
18.3	10	11	0	0
19.3	12	6	0	13
20.3	0	17	7	9

(d) D12S391

a	$z_{1,a}$	$z_{2,a}$	$z_{3,a}$	$z_{4,a}$
14	9	0	0	6
15	10	8	10	8
16	0	8	8	9
17	0	4	0	12
17.3	0	0	10	0
18	45	17	20	58
18.3	0	0	0	0
19	93	82	71	157
19.3	14	8	0	9
20	0	10	0	0
20.3	0	0	0	0
21	11	0	10	10
21.3	0	0	0	0
22	9	0	13	13
23	0	9	13	0
24	0	0	0	6
25	12	0	11	8
26	16	0	0	10
27	11	14	8	9

Bibliography

Applied Biosystems (2009). *ABI Prism® GeneMapper™ Software Version 4.1*. Applied Biosystems® by Life Technologies™.

Applied Biosystems (2012a). *AmpFℓSTR® NGM™ PCR Amplification Kit*. Applied Biosystems® by Life Technologies™.

Applied Biosystems (2012b). *AmpFℓSTR® SGM Plus® PCR Amplification Kit*. Applied Biosystems® by Life Technologies™.

Balding, D. (2013). Evaluation of mixed-source, low-template DNA profiles in forensic science. *Proceedings of the National Academy of Sciences of the United States of America*, 110(30):12241–12246.

Balding, D. J. (2005). *Weight-of-evidence for Forensic DNA Profiles*. Statistics in Practice. Wiley, Chichester, UK.

Bill, M., Gill, P., Curran, J., Clayton, T., Pinchin, R., Healy, M., and Buckleton, J. (2005). PENDULUM – a guideline-based approach to the interpretation of STR mixtures. *Forensic Science International*, 148:181–189.

Bright, J.-A., Huizing, E., Melia, L., and Buckleton, J. (2011). Determination of the variables affecting mixed minifiler™ DNA profiles. *Forensic Science International: Genetics*, 5(5):381 – 385.

Bright, J.-A., McManus, K., Harbison, S., Gill, P., and Buckleton, J. (2012). A comparison of stochastic variation in mixed and unmixed casework and synthetic samples. *Forensic Science International: Genetics*, 6(2):180 – 184.

Budowle, B., Ge, J., Chakraborty, R., Eisenberg, A., Green, R., Mulero, J., Lagace, R., and Hennessy, L. (2011). Population genetic analyses of the NGM STR loci. *International Journal of Legal Medicine*, 125(1):101–109.

Butler, J. M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, Second Edition*. Elsevier ACADEMIC PRESS.

Butler, J. M., Schoske, R., Vallone, P. M., Redman, J. W., and Kline, M. C. (2003). Allele frequencies for 15 autosomal loci on U.S. Caucasian, African American, and Hispanic populations. *Journal of Forensic Science*, 48(4).

Cowell, R. (1997). Sampling without replacement in junction trees. Statistical research paper 15, Faculty of Actuarial Science and Insurance, City University London, London, United Kingdom.

Cowell, R. G. (2009). Validation of an STR peak area model. *Forensic Science International: Genetics*, 3(3):193 – 199.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.

Cowell, R. G., Graversen, T., Lauritzen, S., and Mortera, J. (2015). Analysis of forensic DNA mixtures with artefacts. *Journal of the Royal Statistical Society, series C*. To appear.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2007a). A gamma model for DNA mixture analyses. *Bayesian Analysis*, 2(2):333–348.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2007b). Identification and separation of DNA mixtures using peak area information. *Forensic Science International*, 166(1):28–34.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2011). Probabilistic expert systems for handling artifacts in complex DNA mixtures. *Forensic Science International: Genetics*, 5:202–209.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap and its Applications*. Cambridge University Press.

Dawid, A. P. (1984). Statistical theory. The prequential approach. *Journal of the Royal Statistical Society, Series A*, 147:277–305.

Dawid, A. P. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2:25–36.

Dawid, A. P., Mortera, J., Pascali, V. L., and van Boxel, D. W. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, 29:577–595.

El Barmi, H. and Dykstra, R. L. (1999). Likelihood ratio test against a set of inequality constraints. *Journal of Nonparametric Statistics*, 11(1-3):233–250.

Ghalanos, A. and Theussl, S. (2012). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.12.

Gill, P., Curran, J., and Elliot, K. (2005). A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic acids research*, 33(2):632–643.

Gill, P., Curran, J., Neumann, C., Kirkham, A., Clayton, T., Whitaker, J., and Lambert, J. (2008). Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. *Forensic Science International: Genetics*, 2(2):91 – 103.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. C. Griffin London.

Good, I. J. (1979). Studies in the History of Probability and Statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, 66(2):393–396.

Graversen, T. (2014). *DNAmixtures: Statistical Inference for Mixed Traces of DNA*. R package version 0.1-3, `dnamixtures.r-forge.r-project.org/`.

Graversen, T. and Lauritzen, S. (2013). Estimation of parameters in DNA mixture analysis. *Journal of Applied Statistics*, 40(11):2423–2436.

Graversen, T. and Lauritzen, S. (2014). Computational aspects of DNA mixture analysis. *Statistics and Computing*. DOI: 10.1007/s11222-014-9451-7.

Green, P. J. and Mortera, J. (2009). Sensitivity of inferences in forensic genetics to assumptions about founder genes. *Annals of Applied Statistics*, 3(2):731–763.

Harris, T. (1963). *The Theory of Branching Processes*. Springer-Verlag.

Harris, T. E. (1948). Branching processes. *The Annals of Mathematical Statistics*, 19(4):pp. 474–494.

HUGIN API (2009). *HUGIN API Reference Manual*. Hugin Expert A/S.

Jo, J., Fortin, J.-Y., and Choi, M. Y. (2011). Weibull-type limiting distribution for replicative systems. arXiv:1103.3038.

Konis, K. and Hugin Expert A/S (2014). *RHugin: RHugin*. R package version 7.8-2.

Lalam, N., Jacob, C., and Jagers, P. (2005). Estimation of the PCR efficiency based on a size-dependent modelling of the amplification process. *Comptes Rendus Mathematique*, 341(10):631 – 634.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford, United Kingdom.

Lindley, D. (1977). A problem in forensic science. *Biometrika*, 64(2):207–213.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley series in probability and statistics, 2 edition.

Mitchell, A. A., Tamariz, J., O'Connell, K., Ducasse, N., Budimlija, Z., Prinz, M., and Caragine, T. (2012). Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in. *Forensic Science International: Genetics*, 6(6):749 – 761.

Mortera, J., Dawid, A. P., and Lauritzen, S. L. (2003). Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, 63:191–205.

Nilsson, D. (1998). An efficient algorithm for finding the M most probable configurations in a probabilistic expert system. *Statistics and Computing*, 8:159–73.

Perlin, M. W., Legler, M. M., Spencer, C. E., Smith, J. L., Allan, W. P., Belrose, J. L., and Duceman, B. W. (2011). Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*, 56(6):1430–1447.

Puch-Solis, R., Rodgers, L., Mazumder, A., Pope, S., Evett, I., Curran, J., and Balding, D. (2013). Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. *Forensic Science International: Genetics*, 7(5):555–563.

Seillier-Moiseiwitsch, F. and Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association*, 88:355–359.

Steele, C. D. and Balding, D. J. (2014). Statistical evaluation of forensic DNA profile evidence. *Annual Review of Statistics and Its Application*, 1(1):361–384.

Stolovitzky, G. and Cecchi, G. (1996). Efficiency of DNA replication in the polymerase chain reaction. *Proceedings of the National Academy of Science, USA*, 93:12947–12952.

Sun, F. (1995). The polymerase chain reaction and branching processes. *Journal of Computational Biology*, 2(1):63–86.

Taylor, D., Bright, J.-A., and Buckleton, J. (2013). The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics*, 7(5):516 – 528.

Tvedebrink, T., Eriksen, P. S., Asplund, M., Mogensen, H. S., and Morling, N. (2012a). Allelic drop-out probabilities estimated by logistic regression—further considerations and practical implementation. *Forensic Science International: Genetics*, 6(2):263 – 267.

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2009). Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics*, 3(4):222 – 226.

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2010). Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures. *Applied Statistics*, 59(5):855 – 874.

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2012b). Identifying contributors of DNA mixtures by means of quantitative information of STR typing. *Journal of Computational Biology*, 19(7):887–902.

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2012c). Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. *Forensic Science International: Genetics*, 6(1):97 – 101.

Yannakakis, M. (1981). Computing the minimum fill-in is NP-complete. *SIAM Journal on Algebraic and Discrete Methods*, 2:77–79.

Ye, Y. (1987). *Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming*. PhD thesis, Department of Electrical Engineering, Stanford University.